Joint Ph.D. in Computer Science

Generic Framework for the Multidimensional Processing and Analysis of Social Media Content
A Proxemic Approach

A joint doctoral thesis presented by

Maxime Masson

University of Pau and Pays de l'Adour [1]
UPPA
Dept. of Computer Science
LIUPPA Laboratory
Pau, France

University of the Basque Country [2]
UPV/EHU
Dept. of Computer Languages and Systems
HiTZ Center - Ixa Research Group
Donostia-San Sebastián, Spain

A thesis submitted in fulfillment of the requirements for the degrees of

Doctor in Computer Science [1]

Doctor in Language Analysis and Processing [2]

Supervisors:	Dr. Christian Sallaberry	LIUPPA, E2S, University of Pau and Pays de l'Adour
	Dr. Rodrigo Agerri	HiTZ Center - Ixa, Univ. of the Basque Country
Advisors:	Dr. Marie-Noelle Bessagnet	LIUPPA, E2S, University of Pau and Pays de l'Adour
	Prof. Philippe Roose	LIUPPA, E2S, University of Pau and Pays de l'Adour
	Dr. Annig Le Parc Lacayrelle	LIUPPA, E2S, University of Pau and Pays de l'Adour
Reviewers:	Prof. Josiane Mothe	IRIT, CNRS, INSPE, University of Toulouse
	Prof. Elena Cabrio	I3S, INRIA, CNRS, Côte d'Azur University
Examiners:	D.R. Ana-Maria Olteanu-Raimond	LASTIG, IGN/ENSG, Gustave Eiffel University
	D.R. Maguelonne Teisseire	UMR TETIS, INRAE, University of Montpellier

Jury President: Research Director Ana-Maria Olteanu-Raimond

Submitted **June 10th, 2024** ~ Defended **September 23rd, 2024**

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Christian Sallaberry and Dr. Rodrigo Agerri, and my advisors, Dr. Marie-Noelle Bessagnet, Prof. Philippe Roose, and Dr. Annig Le Parc Lacayrelle. Their kindness, trust, and exceptional guidance throughout my Ph.D. journey, along with their constant availability to answer my questions, organize meetings, and review my documents, were invaluable. I am also grateful for the opportunity to pursue this Ph.D. after my Master's research internship. My thesis journey unfolded under the best possible conditions thanks to their unwavering support, and their expertise allowed me to complete this project. It was not always easy, but we did it together.

I would also like to extend my gratitude to the LIUPPA Laboratory of the University of Pau and Pays de l'Adour (UPPA) and the HiTZ Center, Ixa Research Group of the University of the Basque Country (UPV/EHU) for granting access to their materials and facilities. My thanks also go to E2S UPPA and the Pau Béarn Pyrénées Urban Community for funding my thesis. Additionally, I appreciated the warm welcome from the entire teaching team at the IAE Pau-Bayonne and at the HiTZ Center, both the teachers, students, and the administrative staff, as my offices were located in their building.

A special thank you to my reviewers, Prof. Josiane Mothe and Prof. Elena Cabrio, for accepting the role and taking the time to read and review my work. I am also thankful to my examiners, Research Director Ana-Maria Olteanu-Raimond and Research Director Maguelonne Teisseire, for attending my defense and listening to my presentation.

To my family, my parents and my sister, I am profoundly grateful for their unwavering support and for allowing me to fully dedicate myself to this thesis without having to worry about anything else. They always believed in my success and supported me.

I also want to express my sincere appreciation to Dr. Sébastien Laborie for inviting me to accompany him to Naples, Italy, to present my research at the University of Parthenope. My heartfelt thanks also go to Dr. Asmaa Ata for her kindness and assistance in establishing the qualitative evaluation protocols. I would like to thank the Basque Country and Pau Béarn Pyrénées Tourism Offices for welcoming me and accepting to share their precious feedback on my work. I am grateful to Dr. Landy Rajaonarivo for connecting me with Kyūshū University and assisting in drafting my postdoctoral project proposal for Japan.

Finally, I would like to thank my friend Najah and colleagues at the laboratory: Charlotte, Virginia, Sélim, Cécile, and Siwar, for their fellowship and support throughout this journey.

I dedicate this thesis to my parents and sister.

Abstract

Generic Framework for the Multidimensional Processing and Analysis of Social Media Content

A Proxemic Approach

In recent decades, significant growth and diversification in sources of User-Generated Content (UGC) have been observed. Social media emerges as one of the primary sources of UGC, offering numerous advantages over traditional data sources, such as affordability, vastness, and diversity across various domains of application (for example, *tourism*, *health*, *public policies*). However, the highly unstructured nature of social media posts introduces several challenges. The language diversity and specificity of social media posts, characterized by features such as brevity, frequent grammatical errors, and the use of special characters, combined with the substantial volume and noisy nature of the data, make analyzing social media data a complex endeavour.

This thesis introduces a novel multilingual framework, the APs Framework, designed to streamline the processing and analysis of social media data. This framework is generic in two aspects: it can be applied across various social media platforms and is adaptable to different application domains. The genericity of the application domain is supported by semantic representations of domain knowledge (for example, through thesaurus or ontologies). The APs Framework aims to provide domain-independent insights from social media to non-computer scientists, such as stakeholders in various domains (for example, tourism offices in the tourism domain), thereby enhancing their analytical capabilities. The APs Framework is structured into four phases: *Collect*, *Transform*, *Analyze*, and *Valorize*.

In the *Collect* phase, a generic and iterative methodology for constructing thematic datasets from social media is proposed. This approach seeks to mitigate the challenges of creating accurate and representative datasets amidst the voluminous and noisy nature of social media. The objective is to shift from ad hoc extraction techniques, prevalent in existing studies, to a more systematic, semi-automatic process. This methodology incorporates human feedback at various stages and utilizes both content-based and metadata-based filtering techniques, alongside semantic domain descriptions, to offer a standardized and reusable method for thematic dataset building from social media. The methodology was evaluated both qualitatively and quantitatively through the development of an X/Twitter dataset focused on tourism in the *Basque Country* region.

The *Transform* phase tackles the challenge of converting multilingual, unstructured text data into structured knowledge within a given application domain. It concentrates on three pivotal knowledge extraction tasks: (1) Sentiment Analysis, (2) Named Entity Recognition (NER) for Locations, and (3) Fine-grained Thematic Concept Extraction. Given the scarcity of multilingual training resources in the tourism domain, the process of manually generating a novel annotated training dataset for this domain is detailed. Subsequently, the thesis explores optimal strategies for the multilingual analysis of social media content in tourism, comparing rule-based and deep

learning-based approaches (including fine-tuning and prompting-based few-shot learning with various language models). This exploration aims to identify the minimal number of annotated examples necessary for achieving competitive results across these tasks, leveraging various training techniques and language models. This phase addresses the challenge of minimizing manual annotation efforts without compromising the results' quality, considering the time-consuming and expensive nature of manual data annotation.

In the *Analyze* phase, we hypothesize that adapting the theory of *proxemics*, traditionally applied in physical contexts, to social media could offer a novel approach to crafting meaningful, domain-adaptable indicators for various end-users. The theory is formally redefined, leading to the development of a modular and extensible proxemic data model. This model is capable of representing social media entities and their interactions in a domain-independent manner. Leveraging this model, *ProxMetrics*, a toolkit and formula for generating adaptable indicators from social media is introduced. These indicators, conceptualized as proxemic similarity measures, span multidimensional social media entities, including users, groups, places, themes, and temporal periods. They are highly customizable, allowing for the adjustment of the five proxemic dimensions (Distance, Identity, Location, Movement and Orientation) to address various domain requirements. The toolkit and models underwent qualitative evaluations in collaboration with a local tourism office to model and address various local touristic requirements.

Finally, the *Valorize* phase addresses the challenge of presenting social media indicators and analyses to non-computer scientist users, such as domain stakeholders, in an accessible and domain-independent manner. To this end, *TextBI*, a multimodal generic dashboard, is proposed. This tool is designed to display multidimensional annotations and indicators over volumes of multilingual social media data, focusing on four core dimensions: spatial, temporal, thematic, and personal, while also accommodating additional enrichment data, such as sentiment and engagement. The dashboard offers various visualization modes, including frequency, movement, association and, *proxemics*, combining features from Business Intelligence (interactivity, combined filtering, synchronization of visuals), Geographical Information Systems (spatial view at multiple granularities), and Linguistic Information Visualization tools (text-based analyses). Unlike most existing dashboards, it is generic to operate across different domains, provided the data adheres to the specified data model. The effectiveness of this dashboard was validated in the tourism domain through evaluations conducted by tourism offices, assessing its applicability and relevance.

The framework's twofold genericity (application domain and data source) is demonstrated through the application of each phase in another domain of application: *local public policies*, leveraging data from municipality review platforms.

Résumé

Cadre Générique pour le Traitement et l'Analyse Multidimensionnelle des Réseaux Sociaux

Une Approche Proxémique

Au cours des dernières décennies, nous avons constaté une expansion significative ainsi qu'une diversification des sources de Contenu Généré par les Utilisateurs (CGU). Les plateformes de réseaux sociaux se distinguent comme une des principales sources de ce type de contenu, présentant plusieurs avantages par rapport aux sources traditionnelles de données, tels que l'accessibilité, la quantité et la variété des données disponibles, et ce, dans des domaines d'application aussi riches que variés comme, par exemple, *le tourisme*, *la santé* et *les politiques publiques*. Toutefois, le caractère fortement non structuré des publications sur les réseaux sociaux soulève de nombreux défis. La diversité linguistique et la spécificité des publications, marquées par des éléments tels que la concision, les erreurs grammaticales fréquentes et l'emploi de caractères spéciaux, en plus du volume considérable et du caractère bruité des données, compliquent l'analyse des données issues des réseaux sociaux.

Cette thèse introduit un nouveau cadre de travail (*framework*) multilingue, le framework APs, conçu pour simplifier le traitement et l'analyse des données des réseaux sociaux. Ce framework est générique sur deux aspects : il peut être appliqué à différentes plateformes de réseaux sociaux et est adaptable à différents domaines d'application. La généricité du framework à des domaines d'application variés est permise grâce à l'utilisation de représentations sémantiques des connaissances du domaine (par exemple, à travers des thésaurus ou des ontologies). Le framework APs vise à extraire des connaissances de manière indépendante du domaine à partir des réseaux sociaux pour des utilisateurs finaux non-informaticiens, tels que les parties prenantes dans divers domaines (par exemple, les offices de tourisme dans le domaine du tourisme), enrichissant ainsi leurs processus d'analyse. Le framework est structuré en quatre phases : *Collecte*, *Transformation*, *Analyse*, et *Valorisation*

Dans la phase de *Collecte*, une méthodologie générique et itérative pour la construction de jeux de données thématiques à partir des réseaux sociaux est proposée. Cette méthodologie vise à surmonter les difficultés liées à la création de jeux de données à la fois précis et exhaustifs, dans le contexte volumineux et saturé de bruit des réseaux sociaux. Elle aspire à évoluer, de techniques d'extraction ad hoc, prévalentes dans les travaux existants, vers un processus formel semi-automatique. Cette méthodologie incorpore des boucles de retours humains à divers stades et combine à la fois des techniques de filtrage basées sur le contenu et sur les métadonnées, en lien avec des descriptions sémantiques des domaines. Le but est d'offrir une méthode standardisée et réutilisable pour la construction de jeux de données thématiques à partir des réseaux sociaux. La méthodologie a été évaluée à la fois qualitativement et quantitativement à travers le développement d'un jeu de données X/Twitter axé sur le tourisme au *Pays Basque*.

La phase de *Transformation* aborde le défi de transformer des données textuelles multilingues et non structurées en connaissances structurées applicables à un domaine spécifique. Cette étape se concentre sur trois tâches récurrentes d'extraction de connaissances : (1) l'analyse des sentiments, (2) la reconnaissance d'entités nommées, spécifiquement pour les lieux, et (3) l'extraction fine de concepts thématiques. Face à la pénurie de ressources d'entraînement multilingues, spécialement dans le secteur du tourisme, nous présentons la création d'un nouveau jeu de données d'entraînement annoté pour ce domaine. Ensuite, la thèse examine les stratégies les plus efficaces pour l'analyse multilingue du contenu sur les réseaux sociaux dans le domaine du tourisme, en comparant des méthodes basées sur des règles et l'apprentissage automatique, y compris l'apprentissage par transfert (*fine-tuning*) et l'apprentissage en peu d'exemples (*few-shot*) avec différents modèles de langage. Cette étude comparative cherche à déterminer le nombre minimal d'exemples annotés requis pour obtenir des performances optimales dans ces trois tâches, en utilisant diverses techniques d'entraînement et modèles de langage. L'objectif est de réduire au maximum les efforts d'annotation manuelle (un processus souvent long et chronophage) tout en préservant la qualité des résultats.

Dans la phase d'*Analyse*, nous explorons l'idée d'adapter la théorie de la *proxémique*, initialement conçue pour les interactions dans des espaces physiques, à l'univers des réseaux sociaux. Cette adaptation vise à proposer une nouvelle méthode pour élaborer des indicateurs pertinents et universels pour des utilisateurs finaux dans divers domaines. La théorie est donc redéfinie dans ce nouveau contexte, ce qui conduit à la création d'un modèle de données proxémique, à la fois modulaire et extensible. Ce modèle est conçu pour représenter les entités sur les réseaux sociaux et leurs interactions de manière générique, sans se limiter à un domaine d'application spécifique. Grâce à ce modèle, nous introduisons *ProxMetrics*, un ensemble d'outils et une formule permettant de créer des indicateurs flexibles basés sur les données des réseaux sociaux. Ces indicateurs, exprimés comme des mesures de similarité proxémique, couvrent des entités multidimensionnelles, comme les utilisateurs, les groupes, les lieux, les thèmes et les périodes temporelles. Ils offrent une grande personnalisation, facilitant l'ajustement des cinq dimensions proxémiques (Distance, Identité, Localisation, Mouvement et Orientation) selon les besoins de chaque domaine. L'utilité de cet ensemble d'outils et de modèles a été étudiée qualitativement en collaboration avec un office de tourisme local, démontrant leur capacité à modéliser et à adresser diverses demandes liées au tourisme.

Enfin, la phase de *Valorisation* a pour objectif de rendre les indicateurs calculés précédemment et les analyses issues des réseaux sociaux accessibles à des utilisateurs non spécialisés en informatique, tels que les parties prenantes d'un domaine, de manière facilement compréhensible et adaptable à divers domaines. À cette fin, *TextBI*, un tableau de bord générique et multidimensionnel, est proposé. Cet outil est conçu pour visualiser des annotations et des indicateurs multidimensionnels sur des jeux de données multilingues provenant des réseaux sociaux. Il se focalise sur quatre dimensions génériques d'analyse : spatiale, temporelle, thématique et personnelle. Il permet également l'intégration de données d'enrichissement supplémentaires, comme le sentiment ou l'engagement. Ce tableau de bord propose divers modes de visualisation, incluant les fréquences, les mouvements et les associations ou encore la *proxémique*. *TextBI* associe des caractéristiques de plusieurs types d'outil comme la Business Intelligence (interactivité, filtres multiples, synchronisation des visuels), les systèmes d'information géographique (vue spatiale à multiples granularités) et les outils de visualisation linguistique (analyses basées sur du texte). Contrairement à la plupart des tableaux de bord existants dans divers domaines, il est polyvalent et fonctionne dans différents domaines, à condition que les données utilisées respectent notre modèle de données proxémique. Ce tableau de bord a été évalué dans le domaine du tourisme avec plusieurs offices de tourisme.

La double généricité (domaine d'application et source de données) du framework est ensuite démontrée par l'application de chacune de ses phases dans un autre domaine : les politiques publiques locales, en exploitant des données provenant de sites web d'avis de villes.

Resumen

Marco Genérico para el Procesamiento y Análisis Multidimensional de Contenido de Redes Sociales

Un Enfoque Proxémico

En las últimas décadas, hemos sido testigos de una expansión significativa, así como de una diversificación de las fuentes de Contenido Generado por los Usuarios (CGU). Las plataformas de redes sociales se destacan como una de las principales fuentes de este tipo de contenido, ofreciendo varias ventajas sobre las fuentes tradicionales de datos, tales como la accesibilidad, la cantidad y la variedad de los datos disponibles, en áreas de aplicación tan ricas como variadas, como, por ejemplo, *el turismo*, *la salud* y *las políticas públicas*. Sin embargo, el carácter altamente no estructurado de las publicaciones en las redes sociales plantea numerosos desafíos. La diversidad lingüística y la especificidad de las publicaciones, marcadas por elementos como la concisión, los errores gramaticales frecuentes y el uso de caracteres especiales, además del volumen considerable y el carácter ruidoso de los datos, complican el análisis de los datos provenientes de las redes sociales.

Esta tesis introduce un nuevo marco de trabajo multilingüe, el marco de trabajo APs, diseñado para simplificar el procesamiento y análisis de los datos de las redes sociales. Este marco es genérico en dos aspectos: puede ser aplicado a diferentes plataformas de redes sociales y es adaptable a diferentes áreas de aplicación. La genericidad del marco a diversas áreas de aplicación es posible gracias al uso de representaciones semánticas del conocimiento del dominio (por ejemplo, a través de tesauros u ontologías). El marco de trabajo APs tiene como objetivo extraer conocimientos de manera independiente del dominio a partir de las redes sociales para usuarios finales no expertos en informática, tales como las partes interesadas en diversos campos (por ejemplo, oficinas de turismo en el ámbito del turismo), enriqueciendo así sus procesos de análisis. El marco está estructurado en cuatro fases: *Recolección*, *Transformación*, *Análisis* y *Valorización*.

En la fase de *Recolección*, se propone una metodología genérica e iterativa para la construcción de conjuntos de datos temáticos a partir de las redes sociales. Esta metodología busca superar las dificultades relacionadas con la creación de conjuntos de datos tanto precisos como exhaustivos, en el contexto voluminoso y saturado de ruido de las redes sociales. Aspira a evolucionar de técnicas de extracción ad hoc, prevalentes en los trabajos existentes, hacia un proceso formal semiautomático. Esta metodología incorpora bucles de retroalimentación humana en varios estados y combina técnicas de filtrado basadas en el contenido y en los metadatos, en relación con descripciones semánticas de los dominios. El objetivo es ofrecer un método estandarizado y reutilizable para la construcción de conjuntos de datos temáticos a partir de las redes sociales. La metodología ha sido evaluada tanto cualitativa como cuantitativamente a través del desarrollo de un conjunto de datos X/Twitter centrado en el turismo en el *País Vasco*.

La fase de *Transformación* aborda el desafío de transformar datos textuales multilingües y no estructurados en conocimientos estructurados aplicables a un dominio específico. Esta etapa se

centra en tres tareas recurrentes de extracción de conocimientos: (1) análisis de sentimientos, (2) reconocimiento de entidades nombradas, específicamente para lugares, y (3) extracción detallada de conceptos temáticos. Ante la escasez de recursos de entrenamiento multilingües, especialmente en el sector del turismo, presentamos la creación de un nuevo conjunto de datos de entrenamiento anotado para este dominio. A continuación, la tesis examina las estrategias más efectivas para el análisis multilingüe de contenido en redes sociales en el ámbito del turismo, comparando métodos basados en reglas y aprendizaje automático, incluyendo el ajuste fino y el aprendizaje few-shot con diferentes modelos de lenguaje. Este estudio comparativo busca determinar el número mínimo de ejemplos anotados requeridos para obtener un rendimiento óptimo en estas tres tareas, utilizando diversas técnicas de entrenamiento y modelos de lenguaje. El objetivo es minimizar los esfuerzos de anotación manual (un proceso a menudo largo y tedioso) mientras se preserva la calidad de los resultados.

En la fase de *Análisis*, exploramos la idea de adaptar la teoría de la proxémica, inicialmente concebida para las interacciones en espacios físicos, al universo de las redes sociales. Esta adaptación busca proponer un nuevo método para elaborar indicadores relevantes y universales para usuarios finales en diversos dominios. La teoría se redefine, por tanto, en este nuevo contexto, lo que conduce a la creación de un modelo de datos proxémico, modular y extensible. Este modelo está diseñado para representar las entidades en las redes sociales y sus interacciones de manera genérica, sin limitarse a un dominio de aplicación específico. Mediante este modelo, introducimos *ProxMetrics*, un conjunto de herramientas y una fórmula que permite crear indicadores flexibles basados en datos de redes sociales. Estos indicadores, expresados como medidas de similitud proxémica, abarcan entidades multidimensionales, como los usuarios, grupos, lugares, temas y períodos temporales. Ofrecen una gran personalización, facilitando el ajuste de las cinco dimensiones proxémicas (Distancia, Identidad, Localización, Movimiento y Orientación) según las necesidades de cada dominio. La utilidad de este conjunto de herramientas y modelos se validó cualitativamente en colaboración con una oficina de turismo local, demostrando su capacidad para modelar y abordar diversas solicitudes relacionadas con el turismo.

Finalmente, la fase de *Valorización* aborda el desafío de hacer accesibles los indicadores calculados previamente y los análisis derivados de las redes sociales a usuarios no especializados en informática, tales como las partes interesadas de un dominio, de manera fácilmente comprensible y adaptable a diversos dominios. Con este fin, se propone *TextBI*, un tablero de control genérico y multidimensional. Esta herramienta está diseñada para visualizar anotaciones e indicadores multidimensionales sobre conjuntos de datos multilingües provenientes de las redes sociales. Se centra en cuatro dimensiones genéricas de análisis: espacial, temporal, temática y personal. También permite la integración de datos de enriquecimiento adicionales, como el sentimiento o el compromiso. Este tablero ofrece varios modos de visualización, incluyendo frecuencias, movimientos y asociaciones o proxémica. *TextBI* combina características de varios tipos de herramientas, como la Business Intelligence (interactividad, sincronización de visuales), los sistemas de información geográfica (vista espacial a múltiples granularidades) y las herramientas de visualización del lenguaje (análisis basados en texto). A diferencia de la mayoría de los tableros de control existentes en varios dominios, es versátil y funciona en diferentes dominios, siempre que los datos utilizados se adhieran a nuestro modelo de datos proxémico. Este tablero fue evaluado en el ámbito del turismo con varias oficinas de turismo.

La doble genericidad (área de aplicación y fuente de datos) del marco de trabajo se demuestra luego aplicando cada una de sus fases en otro dominio: *las políticas públicas*, explotando datos provenientes de sitios web de reseñas de ciudades.

Laburpena

Sare Sozialen Analisi eta Prozesamendu Multidimentsionalerako Marko Orokorra

Hurbiltasun-Ikuspegi bat

Azken hamarkadetan, Erabiltzaileek Sortutako Edukiaren (ESE) iturrien hedapen eta dibertsifikazio nabarmena ikusi dugu. Sare sozialen plataformak eduki mota honen iturri nagusietako bat bezala nabarmentzen dira, datuen iturri tradizionalen aldean hainbat abantaila eskainiz, hala nola irisgarritasuna, datuen kopurua eta eskuragarri dauden datuen aniztasuna, turismoa, osasuna eta politika publikoak bezalako aplikazio-eremu aberats eta anitzetan. Hala ere, sare sozialen argitalpenen izaera oso egituratu gabekoak hainbat erronka planteatzen ditu. Argitalpenen aniztasun linguistikoa eta berezitasuna, laburtasuna, maizko gramatika-akatsak eta karaktere berezien erabilera bezalako elementuekin markatuta, datu-bolumen handia eta zarata-maila kontuan hartuta, sare sozialetatik datozen datuen analisia konplikatzen dute.

Tesi honek sare sozialen datuen prozesamendua eta analisia errazteko diseinatutako marko orokor berri bat, *APs Framework*, aurkezten du. Marko hau bi alderditan orokorra da: sare sozialen plataformetara eta aplikazio-eremu desberdinetara aplika daiteke. Aplikazio-eremu anitzetara markoaren orokortasuna, besteak beste, ezagutza-eremuko ordezkaritza semantikoak erabiliz (adibidez, tesaurus edo ontologiak bidez) lortzen da. *APs Framework*ek sare sozialetatik domeinu-independenteko ezagutzak erauztea du helburu, informatikariak ez diren erabiltzaileentzat, hala nola turismo-bulegoetako eragileentzat turismoaren arloan, haien analisi-prozesuak aberastuz. Markoa lau faseetan egituratuta dago: *Bilketa*, *Transformazioa*, *Analisiak*, eta *Balioztatzea*

Bilketa fasean, sare sozialetatik datu multzo tematikoak eraikitzeko metodologia orokor eta iteratibo bat proposatzen da. Metodologia honek sare sozialen testuinguru zaratu eta handian zehatzak eta exhaustiboak diren datu multzoak sortzeko zailtasunei aurre egitea du helburu. Ad hoc ateratze-tekniketatik prozesu formal semi-automatiko batera eboluzionatzea asmo du. Metodologia honek etapa desberdinetan giza-feedback zikloak txertatzen ditu eta edukiaren eta metadatuen oinarritutako iragazketa-teknikak konbinatzen ditu, domeinuen deskribapen semantikoekin lotuta. Datu multzo tematikoak sortzeko metodo estandarizatu eta berrerabilgarri bat eskaintzea da helburua. Metodologia hau kalitatezko eta kantitatezko ebaluazio batetik igaro da, Euskal Herrian turismoari buruzko X/Twitter datu multzo bat garatuz.

Transformazioa faseak hizkuntza anitzeko eta egituratu gabeko testu-datuak domeinu zehatz bateko ezagutza egituratura bihurtzeko erronkari aurre egiten dio. Etapa honek ezagutza ateratzeko hiru zeregin errepikakorretan zentratzen da: (1) sentimenduen analisia, (2) izen bereko entitateen aitorpena, bereziki lekuak, eta (3) kontzeptu tematiko zehatzen ateratzea. Hizkuntza anitzetarako prestakuntza-baliabideen eskasiari aurre eginez, batez ere turismoaren sektorean, turismoaren arloan zehazki etiketatutako prestakuntza-datu multzo berri bat sortzen dugu. Ondoren, tesiak sare sozialetako edukien analisi hizkuntza anitzeko estrategiarik eraginkorrenak aztertzen ditu

turismoaren arloan, arauetan oinarritutako metodoak eta ikaskuntza automatikoa, hala nola *fine-tuning* eta *few-shot* ikaskuntza, hizkuntza eredu desberdinekin konparatuz. Ikerketa konparatibo hau hiru zeregin hauetan emaitza onenak lortzeko beharrezkoak diren adibide etiketatuen gutxieneko kopurua zehazten saiatzen da, entrenamendu-teknika eta hizkuntza-eredu desberdinak erabiliz. Helburua da eskuzko etiketatze-lanak (prozesu luze eta denbora-eragilea sarritan) ahalik eta gehien murriztea, emaitzen kalitatea mantenduz.

Analisi fasean, espazio fisikoetako elkarreraginetarako hasiera batean diseinatutako *proxemika* teoria sare sozialetako unibertsora egokitzea proposatzen dugu. Egokitzapen hau domeinu desberdinetako erabiltzaile amaierentzako adierazle errelevantziaz eta unibertsalez metodologia berri bat garatzeko asmoarekin egiten da. Beraz, teoria testuinguru berri honetan berriz definitzen da, datu proxemiko eredu bat sortuz, modulagarria eta hedagarria dena. Eredu hau sare sozialetako entitateak eta haien elkarreraginak modu generikoan irudikatzeko diseinatuta dago, aplikazio-eremu zehatz bati mugatu gabe. Eredu honi esker, *ProxMetrics* tresna-multzo bat eta sare sozialetako datuetan oinarritutako adierazle malguak sortzeko formula bat sartzen ditugu. Adierazle hauek, proxemika antzekotasun-neurri gisa adierazita, erabiltzaileak, taldeak, lekuak, gaiak eta denbora-aldiak bezalako entitate multidimentsionalak estaltzen dituzte. Pertsonalizazio handia eskaintzen dute, bost proxemika dimentsioak (Distantzia, Identitatea, Kokapena, Mugimendua eta Orientazioa) eremuko beharretara egokituz. Tresna-multzoaren eta ereduen erabilgarritasuna kalitatezko lankidetzan ebaluatu da tokiko turismo-bulego batekin, turismoarekin lotutako eskakizun anitzak modelatzen eta helburutzat dituzten gaitasuna erakutsiz.

Azkenik, *Balioztatze* faseak aurretik kalkulatutako adierazleak eta sare sozialetatik ateratako analisiak informatikariak ez diren erabiltzaileentzako, hala nola domeinuko parte-hartzaileentzat, modu erraz ulergarrian eta domeinu desberdinetara egokitu daitekeen moduan eskuragarri jartzearen erronkari aurre egiten dio. Helburu horrekin, *TextBI*, mahaigaineko tresna generiko eta multidimentsional bat proposatzen da. Tresna hau sare sozialetatik datozen datu multilingueetako anotazioak eta adierazle multidimentsionalak bistaratzeko diseinatuta dago. Lau dimentsio generikoetan arreta jartzen du analisian: espaziala, denborala, tematikoa eta pertsonala. Gainera, sentimendu edo inplikazio moduko datu osagarriak integratzeko aukera eskaintzen du. Mahaigain honek bistaratze-modu desberdinak proposatzen ditu, hala nola maiztasunak, mugimenduak eta elkarteak edo proxemika. *TextBI* Business Intelligence tresnen (interaktibitatea, bistaratzeen sinkronizazioa), informazio geografikoaren sistemak (espazio-ikuspegi anitzeko granularitateak) eta bisualizazio linguistikorako tresnak (testuan oinarritutako analisiak) ezaugarriak biltzen ditu. Beste mahaigain batzuen aldean, domeinu desberdinetan funtzionatu dezake, erabilitako datuek gure datu proxemiko ereduari jarraitzen badiote. Mahaigain hau turismoaren arloan hainbat turismo-bulegorekin ebaluatu da.

Markoaren bikoitza orokortasuna (aplikazio-eremua eta datu-iturria) beste eremu batean erakusten da: politika publikoak, hirien iritzien webguneetako datuak erabiliz, bere fase bakoitza aplikatuz.

Important Information

This thesis **makes extensive use of colors** in *figures*, *charts*, and *text* to convey critical information and enhance the readability and understanding of the content. Viewing the document in black and white, or in grayscale, may result in the loss of important information and nuances. Therefore, it is strongly recommended to view this document in color to fully appreciate the visual elements and to ensure accurate interpretation of the data and concepts presented.

Contents

Publications

Alongside this thesis, I have authored several publications that reflect various aspects of my work. Below is a list of these publications.

International Publications

Published in peer-reviewed international conferences and journals

1. M. Masson, P. Roose, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, R. Agerri. (2024). ProxMetrics: Modular Proxemic Similarity Toolkit to Generate Domain-Adaptable Indicators from Social Media. In *Social Network Analysis And Mining*, 14, 124. Springer (Impact Factor: 2.8).

2. M. Masson, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose, R. Agerri. (2024). TextBI: An Interactive Dashboard for Visualizing Multidimensional NLP Annotations in Social Media Data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2024)* (pp. 1-9) (St. Julians, Malta). Association for Computational Linguistics (CORE Rank: A, ERA Rank: A).

3. M. Masson, P. Roose, C. Sallaberry, R. Agerri, M. N. Bessagnet, A. Le Parc Lacayrelle. (2023). APs: A Proxemic Framework for Social Media Interactions Modeling and Analysis. In *International Symposium on Intelligent Data Analysis (IDA 2023)* (pp. 287-299) (Louvain-La-Neuve, Belgium). Cham: Springer Nature Switzerland (CORE Rank: B, ERA Rank: A).

4. M. Masson, C. Sallaberry, R. Agerri, M. N. Bessagnet, P. Roose, A. Le Parc Lacayrelle. (2022). A Domain-independent Method for Thematic Dataset Building from Social Media: The Case of Tourism on Twitter. In *International Conference on Web Information Systems Engineering (WISE 2022)* (pp. 11-20) (Biarritz, France). Cham: Springer International Publishing (CORE Rank: B, ERA Rank: A).

National Publications

Published in peer-reviewed national journals, conferences, and workshops

1. M. Masson, R. Agerri, C. Sallaberry, M. N. Bessagnet, P. Roose, A. Le Parc Lacayrelle. (2024). Optimal Strategies for the Multidimensional Analysis of Multilingual Content from Social Media. In *Proceedings of the 42nd Conference on Computer Science for Organizations and Information and Decision Systems (INFORSID 2024)* (Nancy, France).

2. M. Masson, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose, R. Agerri. (2024). TextBI: Interactive Visualization of Multidimensional Data from Social Media. In *Mappemonde. Quarterly Journal on the Geographic Image and the Forms of the Territory* (to be published).

3. M. Masson, S. Abdelhedi, C. Sallaberry, R. Agerri, M. N. Bessagnet, A. Le Parc-Lacayrelle, P. Roose. (2023). Interactive Visualization of Tourist Activity Trajectories: Application to Data

Extracted from Twitter. In *Workshop "Exploring traces in an all-digital world: challenges and perspectives" at INFORSID 2023* (La Rochelle, France).

4. M. Masson. (2022). Augmented Proxemic Services for Cultural Heritage and Tourism Practices. In *Young Researchers' Forum at INFORSID 2022* (Dijon, France).

Ongoing Publications

Currently in peer-review at international journals

1. M. Masson, R. Agerri, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose. (2023). Optimal Strategies to Perform Multilingual Analysis of Social Content for a Novel Dataset in the Tourism Domain. Submitted to *Knowledge-Based Systems* journal (Impact Factor: 8.8).
    - **History**: Initial paper submitted on January 8th 2024, revision requested on February 29th 2024, revised paper submitted on May 11th 2024, currently awaiting second review.

Previous Publications

As part of my master's research internship, I have also contributed to the following publications, which, while not directly related to the subject of this thesis, reflect my ongoing commitment and contributions to the field.

1. C. Cayèré, C. Sallaberry, C. Faucher, M. N. Bessagnet, P. Roose, M. Masson. (2024). Semantic Trajectories Similarity: Measures Addressing Spatial,Temporal and Thematic Dimensions. In *Open Journal in Information Systems Engineering*, 4, 2. ISTE Open Science.

2. C. Cayèré, C. Sallaberry, C. Faucher, M. N. Bessagnet, P. Roose, M. Masson. (2022). Similarity Measurement for Semantic Trajectories: Taking into Account Three Levels of Granularity. In *Proceedings of the 40th Conference on Computer Science for Organizations and Information and Decision Systems (INFORSID 2022)* (Dijon, France). Recipient of an Outstanding Paper Award.

3. M. Masson, C. Cayèré, M. N. Bessagnet, C. Sallaberry, P. Roose, C. Faucher. (2022). An ETL-like Platform for the Processing of Mobility Data. In *Proceedings of the 37th ACM Symposium on Applied Computing (ACM SAC 2022)* (pp. 547-555) (Brno, Czech Republic) (CORE Rank: B, ERA Rank: B).

4. M. Masson, C. Cayèré, M. N. Bessagnet, C. Sallaberry, P. Roose, C. Faucher. (2022). Spatio-temporal Visualization of Outdoor Tourist Mobility Data . In *Workshop on Spatial and Temporal Data Management and Analysis (GAST), Francophone Conference on Knowledge Extraction and Management (EGC 2022)* (Blois, France)

5. C. Cayèré, C. Sallaberry, C. Faucher, M. N. Bessagnet, P. Roose, M. Masson, J. Richard. (2021). Multi-level and Multiple Aspect Semantic Trajectory Model: Application to the Tourism Domain. *ISPRS - International Journal of Geo-Information*, 10(9), 592. (Impact Factor: 3.4)

Talks

Presentations given at workshops, seminars, and webinars that do not have published proceedings can be found here.

Workshops

1. M. Masson. (November, 2023). TextBI: A Generic Dashboard for Interactive Visualization of Multidimensional Data from Social Media. *Workshop on Spatialized Digital Humanities, Annual Meeting of the GdR CNRS MAGIS (CNRS Research Network on Methods and Applications for Geomatics and Spatial Information)*. Maison des Suds (Bordeaux, France).

2. M. Masson, S. Laborie. (June, 2023). A Generic Framework for the Extraction, Processing, Analysis and, Valorization of Social Media Content. *Symposium "Constitution of corpus for the needs of digital marketing in the domain of fashion" (European Cassini Program)*, Parthenope University of Naples (Naples, Italy).

3. M. Masson. (November, 2022). APs: A Proxemic Approach for Data Analysis on Social Media. *Workshop Smart city, smart destination: from management to territorial experience*, IRGO - Research Institute in Organizational Management, University of Bordeaux (Bordeaux, France).

4. M. Masson. (September, 2022). APs: A Proxemic Approach for Data Analysis on Social Media. *Inter-association Day EGC/INFORSID*, IRIT, University of Toulouse III (Toulouse, France).

Webinars and Seminars

1. M. Masson. (January, 2024). TextBI: An Interactive Platform for Visualizing Multidimensional Data from Social Media. Invited Speaker: *Webinar on Cartography and Geovisualization of the GdR CNRS MAGIS (CNRS Research Network on Methods and Applications for Geomatics and Spatial Information)* (Online).

2. M. Masson, P. Roose. (July, 2023). Analyzing Touristic Data in the Basque Country. *Urban community of the Basque Country* (Bayonne, France).

3. M. Masson. (June, 2023). A Generic Framework for the Extraction, Processing, Analysis, and Valuation of Social Media content: Application to the Domain of Tourism and the Social Media Twitter. *Ixa Seminar*, University of the Basque Country (EHU/UPV) (San Sebastian, Spain).

Awards

1. Winner of the Geodata Challenge (ranked 1st) at the *National Geonumeric Days 2023* (GeoDataDays 2023) held at the *Reims Convention Center* on September 12-13 2023 with the proposal "*Visualization of Data from Social Media: A Business Intelligence-like Platform*". This event was organized by the French Association for Geographic Information (Afigéo) and DécryptaGéo.

Teaching and Supervision

In addition to my research efforts, I have actively engaged in teaching and supervision activities during my Ph.D., which have significantly enriched my academic experience and perspective. These positions have played a pivotal role in shaping my approach to pedagogy and management. Below is a summary of my teaching and supervision engagements, which complement my journey as a researcher and educator.

Teaching

1. *Development of Web Applications*. September to December 2024. 2nd year of Bachelor in Computer Science (L2), Faculty of Science (STEE), University of Pau and Pays de l'Adour (85h lectures and tutorials). Course coordinator (administrative responsibility).

2. *Operating Systems*. September to December 2024. 1st year of Bachelor in Computer Science (L1), Faculty of Science (STEE), University of Pau and Pays de l'Adour (34h tutorials).

3. *Operating Systems*. October 2023. 1st year of Bachelor of Technology in Computer Science (BUT), University Institute of Technology (IUT) of Bayonne and the Basque Country (9h tutorials).

4. *Server-side Programming*. September to December 2023. 2nd year of Bachelor of Technology in Computer Science (BUT), University Institute of Technology (IUT) of Bayonne and the Basque Country (27h tutorials).

5. *Development of Web Applications*. September to December 2023. 2nd year of Bachelor in Computer Science (L2), Faculty of Science (STEE), University of Pau and Pays de l'Adour (25h lectures and tutorials). Course coordinator (administrative responsibility).

6. *Office Automation*. September to December 2023. Professional Bachelor in Insurance (LP), European Research Center for Family, Insurance, Personal, and Health Law - CERFAPS, University of Bordeaux (23h lectures). Course designer and coordinator (administrative responsibility).

7. *Office Automation*. September to December 2022. Professional Bachelor in Insurance (LP), European Research Center for Family, Insurance, Personal, and Health Law - CERFAPS, University of Bordeaux (23h lectures). Course designer and coordinator (administrative responsibility).

8. *Computer Science I*. September to December 2022. 1st year of Bachelor in Economy and Management (L1), Faculty of Social Sciences and Humanities (SSH), University of Pau and Pays de l'Adour (32h tutorials).

9. *Computer Science II*. January to June 2022. 2nd year of Bachelor in Economy and Management (L2), Faculty of Social Sciences and Humanities (SSH), University of Pau and Pays de l'Adour (32h tutorials).

Supervision

1. Co-supervision of Master's Students (January to June 2024): Alexy Del Amo Alonso and Maxime Silla from the *University of Pau and Pays de l'Adour, France*.
   Master's Project Title: "Extraction of Web Documents About Marine Biology".

2. Co-supervision of an Intern (February to August 2023): Siwar Abdelhedi from the *Higher Institute of Computer Science, Tunisia, ISI*.
   Internship's Project Title: "Generic, Multi-dimensional, and Multi-level Visualization of Data from Social Media: Application to the Tourism Domain".

3. Co-supervision of Master's Students (January to June 2023): Victor Laffarguette and Thomas Procureur from the *University of Pau and Pays de l'Adour, France*.
   Master's Project Title: "Annotation of Tourism-Related Phrases in Tweets".

4. Co-supervision of Master's Students (January to June 2022): Aitor Cachenaut and Benjamin Laby from the *University of Pau and Pays de l'Adour, France*.
   Master's Project Title: "Extraction of Spatial and Thematic Named Entities from Multilingual Tweets".

Other

1. Organization Volunteer (2024) at the 22$^{\text{nd}}$ International Conference on Pervasive Computing and Communications (PerCom 2024) (CORE Rank: A*, ERA Rank: A). https://percom.org/2024/

2. Participation in the 2022 edition of the challenge *"My Thesis in 180 Seconds"*. https://mt180.fr/

List of Figures

List of Tables

List of Acronyms

5W1H	**W**ho, **W**hat, **W**hen, **W**here, **W**hy, **H**ow
ABSA	**A**spect-**B**ased **S**entiment **A**nalysis
AI	**A**rtificial **I**ntelligence
APs	**A**ugmented **P**roxemic **s**ervices
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
BI	**B**usiness **I**ntelligence
B-LOC	**B**eginning **Loc**ation
DILMO	**D**istance, **I**dentity, **L**ocation, **M**ovement, **O**rientation (Proxemic Dimensions)
DMO	**D**estination **M**arketing **O**rganization
DSD	**D**omain **S**pecific **D**ashboard
EntLM	**Ent**ity-oriented **L**anguage **M**odel
FS	**F**ew-**S**hot
FT	**F**ine-**T**uning
GIS	**G**eographic **I**nformation **S**ystem
GoLLIE	**G**uideline-f**o**llowing **L**arge **L**anguage Model for **I**nformation **E**xtraction
GPT	**G**enerative **P**re-trained **T**ransformer
IE	**I**nformation **E**xtraction
I-LOC	**I**ntermediate **Loc**ation
LLaMA	**L**arge **La**nguage **M**odel Meta **A**I
LLM	**L**arge **L**anguage **M**odel
LM	**L**anguage **M**odel
LOC	**Loc**ation
LV	**L**inguistic Information **V**isualization
MLM	**M**asked **L**anguage **M**odel
mBert	**M**ultilingual **BERT**
NER	**N**amed **E**ntity **R**ecognition
NLP	**N**atural **L**anguage **P**rocessing
NLU	**N**atural **L**anguage **U**nderstanding
NLG	**N**atural **L**anguage **G**eneration
OECD	**O**rganisation for **E**conomic **C**o-operation and **D**evelopment
OSM	**O**pen **S**treet **M**ap
OTA	**O**nline **T**ravel **A**gency
PET	**P**attern-**E**xploiting **T**raining
POI	**P**oint **o**f **I**nterest
SA	**S**entiment **A**nalysis
SF	**S**et**F**it
UGC	**U**ser-**G**enerated **C**ontent
VR	**V**irtual **R**eality
WTO	**W**orld **T**ourism **O**rganization
XLM-R	**XLM**-**R**oBERTa
XLM-T	**XLM**-RoBERTa fine-tuned for **T**witter (X)

Chapter 1

Introduction

"We shape our tools and thereafter our tools shape us."
— Marshall McLuhan, Canadian Philosopher

In this first chapter, we begin by outlining the general context of the thesis, focusing on our data sources of interest: user-generated content (Section 1.1) and social media (Section 1.2). We discuss the increasing importance of these sources for analyzing and understanding behaviors and phenomena across various application domains, along with the associated difficulties. To illustrate this point, we introduce a motivating scenario in the tourism domain (Section 1.3). Next, we outline the core objectives of the thesis and describe the associated research challenges and hypotheses (Section 1.4). We then briefly highlight our contributions and the publications that validate them (Section 1.5), positioning these contributions within a generic social media processing and analysis framework. Finally, we provide an overview of the organization of this manuscript (Section 1.6).

1.1 Web 2.0 and the Rise of User-Generated Content

In recent decades, we have witnessed significant growth and diversification in sources of user-generated data. User-Generated Content (UGC), which comes from individuals who voluntarily contribute to the community (Krumm et al., 2008), can take various forms, such as *text, pictures, audio, video*, or *interactive content*, originating from a wide variety of sources. With the advent of Web 2.0[1] (O'reilly, 2009), the web has become more interactive and collaborative, hosting these diverse sources. They range from traditional social media platforms, like *X/Twitter* and *Facebook*, to more specialized ones, such as *Foursquare*, review sites like *TripAdvisor* and *Google Reviews*, as well as discussion forums, blogs, and video sharing websites. Today, internet users do not only consume but also produce and actively share their content. Similarly, platform operators no longer create content themselves but provide the tools for users to do so (Naab and Sehl, 2017). In this thesis, we focus on text-based user-generated content, which continues to be the most common form of user-generated content to date.

User-generated content has become highly beneficial for both researchers and businesses (Moens et al., 2014), acting as a major source of data for analyzing online trends and behaviors. This wealth of information is crucial for understanding consumer preferences, market trends, and

[1] The second generation of the World Wide Web, characterized by greater user interactivity, transforming the Web from a collection of static websites into a dynamic and interactive platform.

broader societal issues. By analyzing user-generated content, organizations can gain immediate feedback and insights, monitor shifts in public sentiment, and identify new trends in online interaction. This information is valuable for making informed strategic decisions, developing products, and crafting marketing strategies, providing an advantage in rapidly changing markets. For researchers, user-generated content offers a unique perspective for studying human behavior in the digital age, contributing to various academic disciplines, including social sciences (Han et al., 2018). Overall, leveraging user-generated content is becoming key to generating innovative insights.

1.2 The Case of Social Media

Social media, also known as *social networks*, has been one of the most prominent sources of user-generated content over the past decades. There are many definitions of social media, but a widely accepted one is the following, from the dictionary Merriam-Webster[2]:

> *"Forms of electronic communication through which users create online communities to share information, ideas, personal messages, and other content."*

To illustrate the growing influence of social media, in 2018, *Facebook* had more than 2 billion active users. Nearly half (48.3%) of the world's population was using social media in 2020, a figure expected to climb to 56.7% by 2025 (Rogers et al., 2021). This includes about 89% of young people in OECD countries[3] who were engaged in social networking online (Roser et al., 2015). The volume of content generated daily on social media platforms is also growing rapidly, with around 500 million *X* tweets, 216 million *Facebook* messages, and 500 million *Instagram* stories issued daily (Rogers et al., 2021; Statista, 2024).

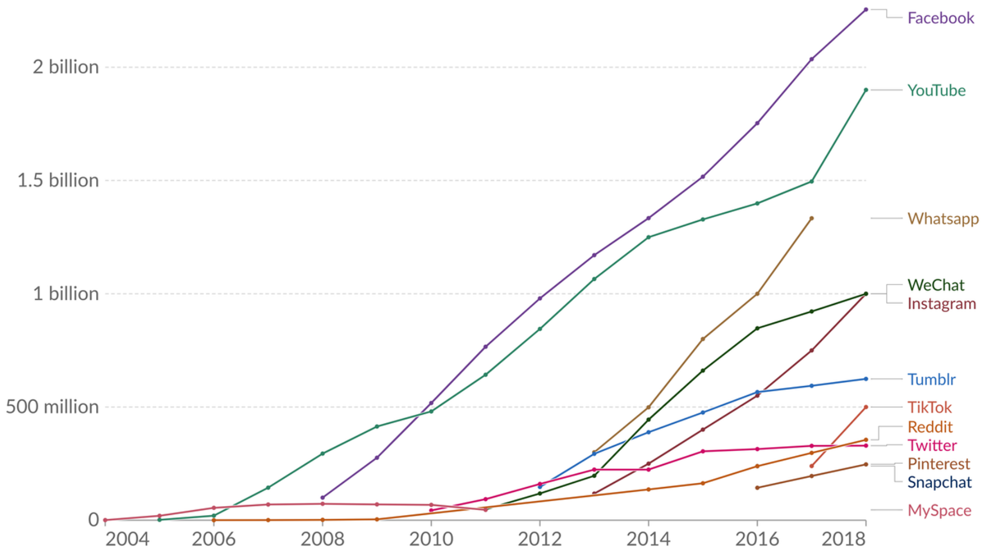


Figure 1.1: Number of People Using Social Media Platforms, 2004 to 2018 – Source: *Our World In Data, CC BY*

[2] https://www.merriam-webster.com

[3] Organization for Economic Co-operation and Development

Social media have thus become an essential resource for analyzing behaviors across a wide variety of topics. As shown in Figure 1.1, the number of social media users has grown exponentially in the last two decades, making it one of the largest sources of user-generated content. In this thesis, our primary focus will be on user-generated content sourced from social media platforms, though our work could apply to other types of user-generated content too.

The exponential growth in social media use, as shown above, not only highlights their significance in modern communication but also underscores their important role in gaining insights into people's behaviors. Let's now highlight the various advantages offered by user-generated content on these platforms for both scientists and businesses. Indeed, user-generated content from social media presents numerous advantages over traditional data sources:

- *Ease of access* and *affordability*: the relative ease of access significantly reduces reliance on expensive commercial datasets or the requirement for time-consuming data collection efforts. Much of this content is available at reduced costs and can be readily accessed, eliminating the barrier of high expenses traditionally associated with acquiring data for analysis.

- *Diversity*: the diversity of user-generated data from social media spans a wide array of application domains, including but not limited to *tourism*, *politics*, and *fashion*. This variety ensures that researchers and businesses can find relevant data across multiple domains of interest, enhancing the breadth and depth of their analyses.

- *Massiveness*: social media platforms usually contain massive amounts of data. This vast quantity of data, while beneficial, introduces its own set of difficulties. Managing and processing enormous volumes of data to extract meaningful insights requires sophisticated tools and techniques.

- *Freshness*: social media platforms are predominantly instantaneous and live, with low latency. New data is posted constantly by users, ensuring that the data remains mostly up-to-date.

Thanks to these numerous advantages, social media have emerged as pivotal tools for gaining insights across a vast array of application domains. Now, let's explore some examples to underscore the critical role of social media data sources in our perpetually connected world.

1.2.1 Impact of Social Media Analyses in Various Application Domains

As mentioned previously, one of the advantages of social media data is their diversity, allowing them to be leveraged to analyze a wide variety of application domains. Additionally, user-generated content on social media is mostly multimodal, comprising both textual content and multimedia, which enables its use in various applications. Let's take an overview of how social media data has been used in four different domains of application over the last decade.

Marketing and Consumer Research

In *marketing and consumer research*, social media platforms have been used for understanding consumer engagement (Barger et al., 2016), brand perception (Yu and Yuan, 2019), or emerging trends for advertising purpose (Wright et al., 2010). Companies use social media analytics to monitor brand mentions, sentiment analysis, and consumer feedback, allowing for more targeted and effective marketing strategies (Alalwan et al., 2017). The live aspect of social media enables companies to quickly identify and respond to consumer trends, adjusting their approaches in real-time to stay competitive in the market (Dey et al., 2011).

Public Health and Epidemiology

In *public health and epidemiology*, social media have become a crucial tool, especially evident during the Covid-19 pandemic (Goel and Gupta, 2020; Tsao et al., 2021). Health organizations and researchers use social media to track disease spread (Sadilek et al., 2012), misinformation (Suarez-Lledo and Alvarez-Galvez, 2021), and public sentiment towards public health measures and vaccinations (Venegas-Vera et al., 2020). This real-time data assists in predicting potential outbreaks, understanding public concerns, and disseminating accurate health information to combat misinformation.

Political Science and Public Opinion

Social media are a powerful source of data for political scientists and policymakers to gauge public opinion (McGregor, 2019), political sentiment (Mejova et al., 2013), and the popularity of policies or political figures (Gainous and Wagner, 2014). Through the analysis of social media data, researchers can identify trends in political discourse (Garimella et al., 2018), public engagement with political events such as elections (Chauhan et al., 2021), and the spread of political ideologies. This is valuable for political campaign strategies (Bello et al., 2019), public policies' formulation (Azzone, 2018), and understanding electoral dynamics (Anstead and O'Loughlin, 2015).

Environmental Science

Environmental scientists use social media to monitor changes in the environment and wildlife (Bergman et al., 2022), often relying on crowdsourced data (Walker et al., 2019). Social media posts can provide early warnings of environmental hazards, such as wildfires (Slavkovikj et al., 2014) or pollution (Zheng et al., 2019), and contribute to biodiversity monitoring (Di Minin et al., 2015) through citizen science projects. This approach allows for a wider geographic coverage and engagement with the public in environmental conservation efforts.

The domain of *tourism* also heavily leverages social media data. As it is the main application of our work, we will go into more detail on this domain later (refer to Subsection 1.3.2).

Here, we have highlighted the extensive role of social media across various application domains, illustrating the diverse insights that can be gleaned from social media data. However, the volume, complexity, and unstructured nature of user-generated content pose significant challenges in processing and analyzing this data, leading us to the next discussion on the obstacles encountered in processing and analyzing social media data.

1.2.2 Main Difficulties in Social Media Processing and Analytics

As observed, social media platforms serve as a critical source of insights for various domain-specific applications, ranging from marketing to public health surveillance and political science. However, the analysis of vast social media datasets comes with various challenges. The unique nature of social media content presents multiple difficulties that researchers and analysts must overcome to extract meaningful information. It is important to note that in this work, our focus will be on text-based content and associated metadata, excluding multimedia content from our analysis. We have identified several key challenges when handling text-based social media data.

Extracting Knowledge from Unstructured Data

One of the primary difficulties in social media analytics arises from the coexistence of two inherently different types of data on these platforms: *structured* and *unstructured* data (Baars and Kemper, 2008). Structured data typically comprises post or profile metadata, such as geotags, author information, and engagement metrics. In contrast, unstructured data mostly consists of the content of the posts themselves, often in the form of raw text (Rout et al., 2018). Unlike the organized information obtained through structured surveys or databases, social media posts, predominantly text-based, lack a consistent format across platforms (Gundecha and Liu, 2012). This lack of uniformity presents a significant obstacle to analysis, as it prevents the content from being seamlessly integrated into predefined analytical processes or models, for example, business intelligence platforms (Wittwer et al., 2017) or geographic information systems (Sui and Goodchild, 2011).

Brevity of Text and Informal Language

Social media posts are usually brief, which can significantly limit the amount of context and detail available for analysis (Jiang et al., 2022). This conciseness often results in ambiguity, as the posts may lack sufficient information to accurately determine the user's intent, sentiment, or the full scope of the message being conveyed. Additionally, the brevity of social media posts often limits the context available for analysis, making it difficult to interpret without additional contextual information. The informal language prevalent in social media, including slang, colloquialisms, and abbreviations, adds another layer of complexity to text analysis (Maylawati et al., 2018). This informal use of language, which varies widely between communities, presents a significant challenge for developing universal tools for Sentiment Analysis or Content Categorization that rely on standard language processing (Lo et al., 2017).

Spelling and Punctuation Errors, Special Characters

Social media's richness in non-textual elements such as emojis (Hasyim, 2019), hashtags (Gerrard, 2018), and URLs (Cao and Caverlee, 2015) introduces unique interpretative challenges. These elements, while significant in conveying meaning or sentiment, are difficult to process using traditional text analysis tools. Emojis, for example, express a wide range of emotions and reactions not easily captured through text alone, and hashtags can provide insights into trends but may also be ambiguous or used ironically (Reyes et al., 2012). The casual nature of social media writing often leads to posts that are rife with spelling and punctuation errors (Clark and Araki, 2011). This disregard for grammatical norms complicates Natural Language Processing (NLP) techniques (Ashraf et al., 2021), which typically rely on correct spelling and grammar to parse and understand text accurately.

Language Diversity (Multilingualism)

The multilingual nature of social media, with users contributing content in various languages and dialects from around the world, requires the use of sophisticated, multilingual NLP tools (Agüero-Torales et al., 2021). This diversity poses significant barriers to comprehensive analysis, as analysts must be prepared to process and understand a multitude of languages.

Data Volume and Diversity

The huge volume of social media posts generated daily presents a daunting challenge in filtering through the noise to isolate relevant data (Stieglitz et al., 2018). The vast amount of content includes a significant proportion of irrelevant or low-quality posts (Jiang et al., 2022), necessitating efficient filtering mechanisms to identify data of interest. Furthermore, social media posts encompass a wide range of topics.

1.3 Thesis Context

We will now explore how this thesis is positioned within the field of social media analysis, and identify the specific aspects it aims to address.

1.3.1 The APs Project

This thesis is part of the APs Project. The APs Project (standing for *Augmented Proxemic services*) is a cross-border French-Spanish project funded by the urban community of *Pau Béarn Pyrénées*[4] and E2S (Energy Environment Solutions) UPPA[5]. This project aims to develop a suite of tools dedicated to processing and analyzing social media data within a specific application domain. The domain of interest should be semantically defined, for example, through a dictionary, a thesaurus, or an ontology. The idea is to equip decision-makers and stakeholders across various domains with a unified set of tools that leverage data from social media to construct new indicators and insights tailored to their specific domain requirements, helping them in the decision-making process. These social media-based tools should complement and enrich their current analysis, not replace them. The genericity of the toolkit must be twofold: (1) genericity across different social media platforms and (2) genericity across various application domains (for example: *tourism, education, healthcare, etc.*).

As an example, let's have a look at Figure 1.2. This figure illustrates the core objectives of the APs Project and the bullet numbers in the following list are the same as the ones in Figure 1.2.

1. *Library of Semantically-defined Application Domains*: This section displays examples of application domains (such as *Tourism*, *Education*, *Health*, etc.), for which the system should provide insights. A fundamental aspect of the project is that these domains are defined through vocabularies contained within semantic resources. These resources can range from simple structures like *dictionaries* to more complex, hierarchical ones such as *thesauruses* or *ontologies*.

2. *Active Application Domain*: This represents the domain of interest currently in use. The domains mentioned earlier are illustrative; theoretically, any domain defined as previously described should be compatible with the project. The system is intended to be generic and capable of working across various domains. Lastly, it is important to note that within the framework of this project, we do not aim to analyze various domains simultaneously. The objective is to enable the reuse of the same system for different domains at different times.

3. *Toolbox* (*Set of Generic Tools to Process and Analyze Social Media Data*): This is the core unit of the project. It represents a hypothetical set of tools that can process and analyze data from

[4] https://www.pau.fr
[5] https://e2s-uppa.eu/en/index.html

social media platforms to produce insights and indicators according to the domain of interest previously chosen. It is a *gray box*, the inner workings of these processes are not fully known to the end users.

4. *Indicators* and *Visualizations*: The results from the *Toolbox* are expressed as indicators and visualizations. This means the raw data from analyses is processed into a form that is easily understandable and useful for end users, who are not necessarily computer scientists.

5. *End Users* (*non-computer scientists*): The final output is designed for end users who are not necessarily computer scientists. The insights and visualizations provided are actionable and accessible for decision-making or further analysis by professionals in the application domains, regardless of their computing expertise.

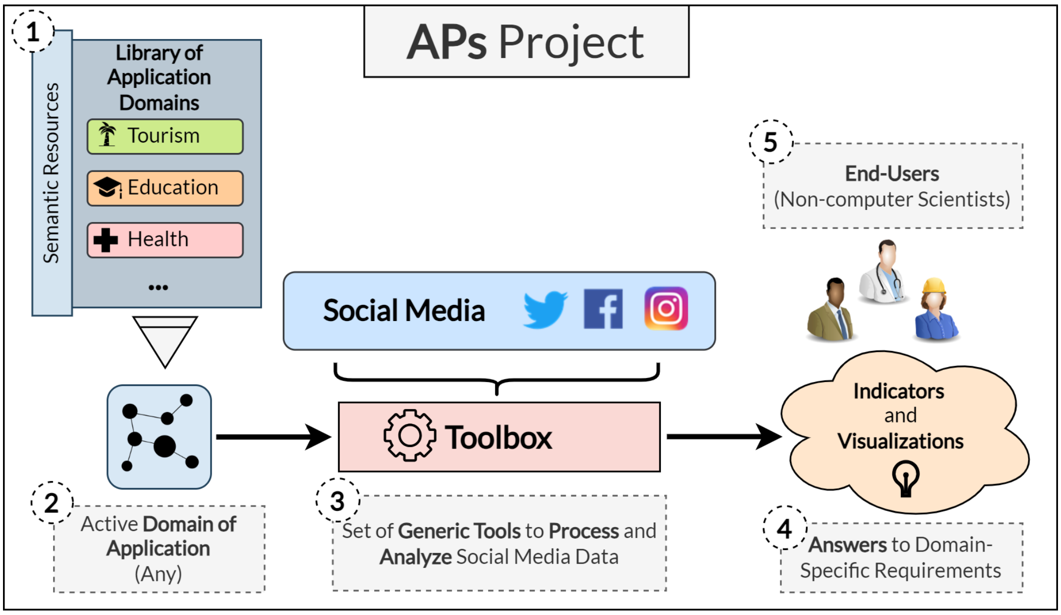

APs Project
1
Semantic Resources
Library of Application Domains
Tourism
Education
Health
...
5
End-Users
(Non-computer Scientists)
Social Media
Toolbox
Indicators and Visualizations
2
Active Domain of Application (Any)
3
Set of Generic Tools to Process and Analyze Social Media Data
4
Answers to Domain-Specific Requirements

Figure 1.2: Overview of the APs Project

It is important to note that the APs Project does not aim to fully replace the rich array of existing analysis techniques in various domains but rather to enrich and complement them by taking into account social media data as well.

The focus areas of this project are the *Basque Country* and *Béarn* regions (highly touristic and multilingual regions spanning across both *France* and *Spain*). Let's now explore how such a system would work and what insights it could provide by taking the domain of tourism in the *Basque Country* region as a case study.

1.3.2 Motivating Scenario – Social Media in the Domain of Tourism

The main application domain of the APs Project is the *domain of tourism*. Tourism is a major economic development lever for the *Basque Country* area. The estimated economic impact of tourism is about 1.6 billion euros, representing nearly 76% of the total tourism economic impact for the *Pyrénées-Atlantiques* department, estimated at around 2 billion euros (Communauté Pays Basque, 2021). However, the tools produced should go further, be generic, and therefore adaptable to any other domain of application that can be defined by a vocabulary (*dictionary, thesaurus, ontology*). In the domain of tourism, social media data can have many practical use cases.

- They can help the decision-making process of tourism stakeholders (Leung et al., 2013) for the improvement, development, and planning of touristic municipalities and areas (Floris et al., 2014). This involves analyzing the data to better understand the practices and requirements of visitors. Such analysis is useful for companies specialized in tourism marketing, such as *Destination Marketing Organizations* (Pike and Page, 2014), where understanding the desires and expectations of visitors is important.

- Touristic data from social media can also be analyzed for visitors themselves by building recommender systems (Majid et al., 2013). These systems analyze the practices of numerous visitors to recommend better-suited places, activities, or touristic itineraries. They can also be used to build user-connection systems to connect visitors with common or divergent interests.

In addition to social media, various sources can be used to extract touristic data. Historically, these primarily include databases in two notable categories: (1) commercial databases, such as those from *Online Travel Agencies* (OTAs) (Suzuki, 2020) or telecom companies, and (2) public databases, for example, those that are government-issued or crowd-sourced (e.g., *DataTourisme* (Boudaa et al., 2021)). Tourism agencies also conduct their own qualitative or quantitative field surveys.

Category	**Description of Indicators and Insights Required**
User Demographics	*Demographics of visitors: age, socio-professional category, etc.*
Activity Combinations	*Before-and-after activities, activity associations.*
Shared and Soft Mobilities	*Indicators on the use of public transport and soft mobility options.*
Cross-Border Travel	*Who? At which locations? Through which medium of transportation?*
Touristic Topics	*Where (hotspots)? What activities? What mobility options?*
Hiking Spaces	*Attendance of hiking areas.*
Swimming Locations	*Attendance of swimming locations (beaches, rivers, lakes, etc.).*
Markets & Farms	*Attendance of markets, farms, etc.*
Accommodations	*Change in accommodations during the stay.*
Camping & Vans	*Parking locations, activities nearby.*
Popular Events	*Major gatherings that generate engagement.*
Business Tourism	*Locations and themes or activities related to professional tourism.*
Weather	*Choice of touristic activities based on weather conditions.*
Visitor Ratio by Topic	*Differentiation between locals, near-locals, close visitors, and others.*
Satisfaction	*Emotions expressed about topics, such as overcrowding or mobility.*
User Reactions	*User reactions to other users' suggestions.*
Influencers	*Identifying key tourism influencers in the Basque Country area.*

Table 1.1: Indicators' Requirements from Social Media of a Local Tourism Office

In this case study, we will focus on the requirements of tourism stakeholders, specifically tourism offices. We met with representatives of the *Tourism Office of the Basque Country*[6] in *Bayonne* and gathered their requirements regarding the analysis of social media data. Currently, they primarily use costly geolocation datasets from telecom companies (e.g., the *Flux Vision*[7] service from *Orange*), and all analyses of social media are conducted manually due to the numerous challenges associated with handling this type of data (refer to Subsection 1.2.2).

Table 1.1 displays an excerpt from requirements collected from this tourism office. Note that (1) we requested that they articulate all aspects they wish to analyze from social media, without limiting their ideas to what they believe is feasible, and (2) tourism offices refer to all people engaging in touristic activities as *visitors* (regardless of whether they are locals or coming from elsewhere).

Let's illustrate how social media, and more specifically the platform X (formerly known as *Twitter*), can be used to address some of these requirements. Below is an example of a tourism-related tweet extracted from X/Twitter, along with its associated metadata (see Figure 1.3). This tweet has been manually annotated across various dimensions, including spatial, temporal, thematic, sentimental, engagement, and personal dimensions.

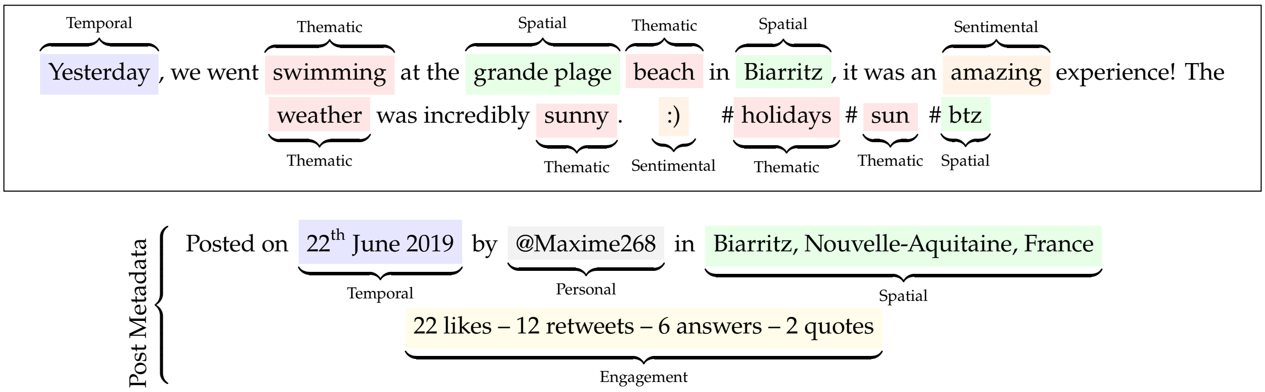

Yesterday, we went swimming at the grande plage beach in Biarritz, it was an amazing experience! The weather was incredibly sunny. :) # holidays # sun # btz
Temporal
Thematic
Spatial
Thematic
Spatial
Sentimental
Thematic
Thematic
Sentimental
Thematic
Thematic
Spatial
Post Metadata
Posted on 22th June 2019 by @Maxime268 in Biarritz, Nouvelle-Aquitaine, France
Temporal
Personal
Spatial
22 likes – 12 retweets – 6 answers – 2 quotes
Engagement

Figure 1.3: Example of Various Dimensions Captured in a Social Media Post

Using large amounts of tweets, we can construct multidimensional trajectories across various dimensions, including personal trajectories of individual users and collective trajectories of groups to address various requirements. Figure 1.4 displays a thematic map (formatted as a *treemap* (Scheibel et al., 2020)) of the domain of tourism as defined by the *Thesaurus on Tourism and Leisure Activities* of the World Tourism Organization (WTO) (World Tourism Organization, 2002) (refer to Appendix A for an extract of the *Accommodation* branch). This thesaurus is an extensive, multilingual glossary in three languages (French, English, and Spanish) covering around 1,300 touristic concepts organized into a dozen main branches (such as *Leisure Activities*, *Economy of Tourism*, *Hospitality*, etc.). In Figure 1.4, the size of each square represents the frequency with which a given touristic concept was mentioned in tweets originating from the *Basque Country* region.

As we can observe, at first glance, social media data provides us with interesting insights into which touristic topics are most prevalent in the region. Over the thematic map, the thematic trajectory of a single visitor is superimposed, giving us insights into which topics or touristic

[6] https://www.en-pays-basque.fr
[7] https://www.orange-business.com/fr/solutions/data-intelligence-iot/flux-vision

activities the users engaged with. This type of trajectory is unique in its *ubiquity*, as a given user can be in several thematic spaces simultaneously. With thousands of trajectories, we could conduct more advanced analyses to detect affinities between types of touristic activities, categories of cultural heritage, or even recurring sequences of activities. These thematic trajectories could also be correlated with other dimensions such as spatial, temporal, thematic, and sentimental, etc., allowing us to address various touristic requirements and, by using a different semantic resource, requirements from other domains.

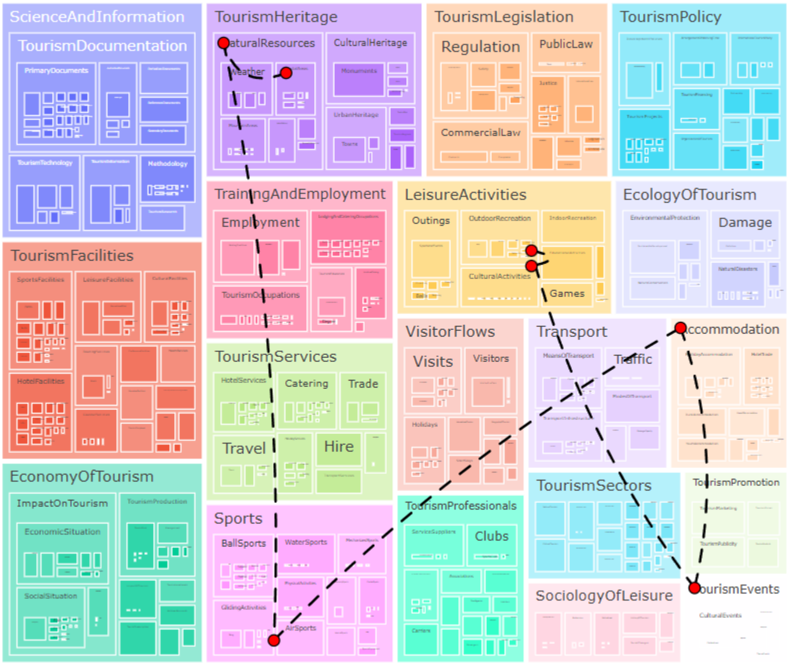


Figure 1.4: Trajectory of a Visitor in What Could be the Tourism Thematic Space

We will now discuss the main proposal of this thesis: a generic and adaptable framework for processing and analyzing data from social media around a chosen domain of interest: the APs Framework. This framework serves as the *toolbox* referenced in Figure 1.2, enabling domain stakeholders to obtain valuable insights around a given domain based on social media.

1.4 Proposal

We propose a generic and adaptable framework for processing and analyzing social media data within any domain of interest, named the APs Framework. This framework is generic in two significant ways:

- *Domain genericity*: It supports any domain of application that can be defined by a semantic resource, be it hierarchized or not, using a specific vocabulary.
- *Source genericity*: It is designed to be compatible with any social media platform that operates on a text-based post system, which is the case for most social media platforms.

The objective of the APs Framework is to produce meaningful indicators and insights from a given domain of interest (which can be any) and a social media source, to address domain-specific requirements. The framework comprises four successive phases (*Collect*, *Transform*, *Analyze*, and *Valorize*), each relying on a common data model. The scientific contributions of this thesis are reflected in these phases. Before going further, it is important to define the concept of *proxemics*, as it is a core element of our framework and is related to the main hypothesis of this thesis.

1.4.1 Background: The *Proxemics* Theory

Proxemics was introduced in the seminal work of the American anthropologist Edward T. Hall (Hall, 1966). He defines *proxemics* as "*the science that studies the organization of space and the effect of distance on interpersonal relations*". Hall studied physical distance and the way it affects and regulates interactions between people. He then went further and linked the concept of distance to *proxemic zones* (Hall et al., 1968). There are four core proxemic zones: (1) the intimate zone (*0 to 0.45 m*) which is mainly used for close physical contact, (2) the personal zone (*0.45 to 1.2 m*) for interactions with very close people such as family or friends, (3) the social zone (*1.2 to 3.6 m*) for regular conversations with strangers, and finally (4) the public zone (*more than 3.6 m*) which is used when speaking to an audience or gathering. It is crucial to note that cultural, social, and physical factors can affect the definition of proxemic zones.

In 2011, Greenberg *et al.* extended Hall's definition of *proxemics* to introduce the notion of proxemic dimensions (Greenberg et al., 2011) (also referred to as DILMO dimensions). They identified five dimensions that can be used to express *proxemics* (Distance, Identity, Location, Movement and Orientation, DILMO). *Proxemics* has been used in various domains to analyze physical interactions. We will go into details later in Chapter 4.

Instead, let's delve into the APs Framework, which extends and leverages this theory to produce indicators that address domain-specific requirements.

1.4.2 The APs Framework

Figure 1.5 illustrates the life cycle of the APs Framework, depicted as a circular processing pipeline. This representation is inspired by the WaterWheel (Bucher et al., 2021). The APs Framework consists of four consecutive phases: (1) *Collect*, (2) *Transform*, (3) *Analyze*, and (4) *Valorize*. Each phase leverages a common data model: the APs Trajectory Model.

Collect covers the entire process of finding and extracting relevant data from social media. It aims to produce a dataset of social media posts based on a specific dataset definition, conducted by the end user in collaboration with computer scientists. The dataset of social media users and posts collected in this phase is raw and will be enriched later. The end user is also involved in evaluating and ensuring the dataset's relevance to their analytical requirements. This phase encompasses dataset definition, extraction, and filtering of posts and their associated metadata, as well as their preview and evaluation.

The second phase, *Transform*, involves applying various modifications and enrichments to the previously collected dataset. This phase aims to extract structured information from unstructured text, such as sentiments, locations, and fine-grained thematic concepts linked to a domain-specific knowledge base (e.g., dictionary, ontology, thesaurus), to better characterize the posts. A battery of NLP modules is applied to the collected posts. Lastly, posts are linked across their various dimensions to construct multidimensional trajectories (e.g., trajectories built from social media posts', encompassing multiple dimensions, including spatial, temporal, thematic, sentimental, etc.).

The third phase, *Analyze*, leverages the previously collected and enriched dataset to compute proxemic metrics (called *proxemic similarity measures*), which are indicators calculated based on the domain of interest requirements. These indicators are used to extract knowledge from the multidimensional social media trajectories built in the previous phase. *Proxemics*, the study of the effect of space and distances on social interactions (Hall et al., 1968), serves as the basis for calculating these indicators (refer to Subsection 1.4.1).

Lastly, *Valorize* enables the visualization of results from the previous analysis for end users who are not computer scientists, such as tourism stakeholders. For tourism professionals, multidimensional dashboards and maps could visualize trends and associations of themes and places in social media. We also envisage using the indicators produced as inputs for a tourism recommender system, including activities, places of interest, and itineraries, or a system to connect visitors sharing common interests.

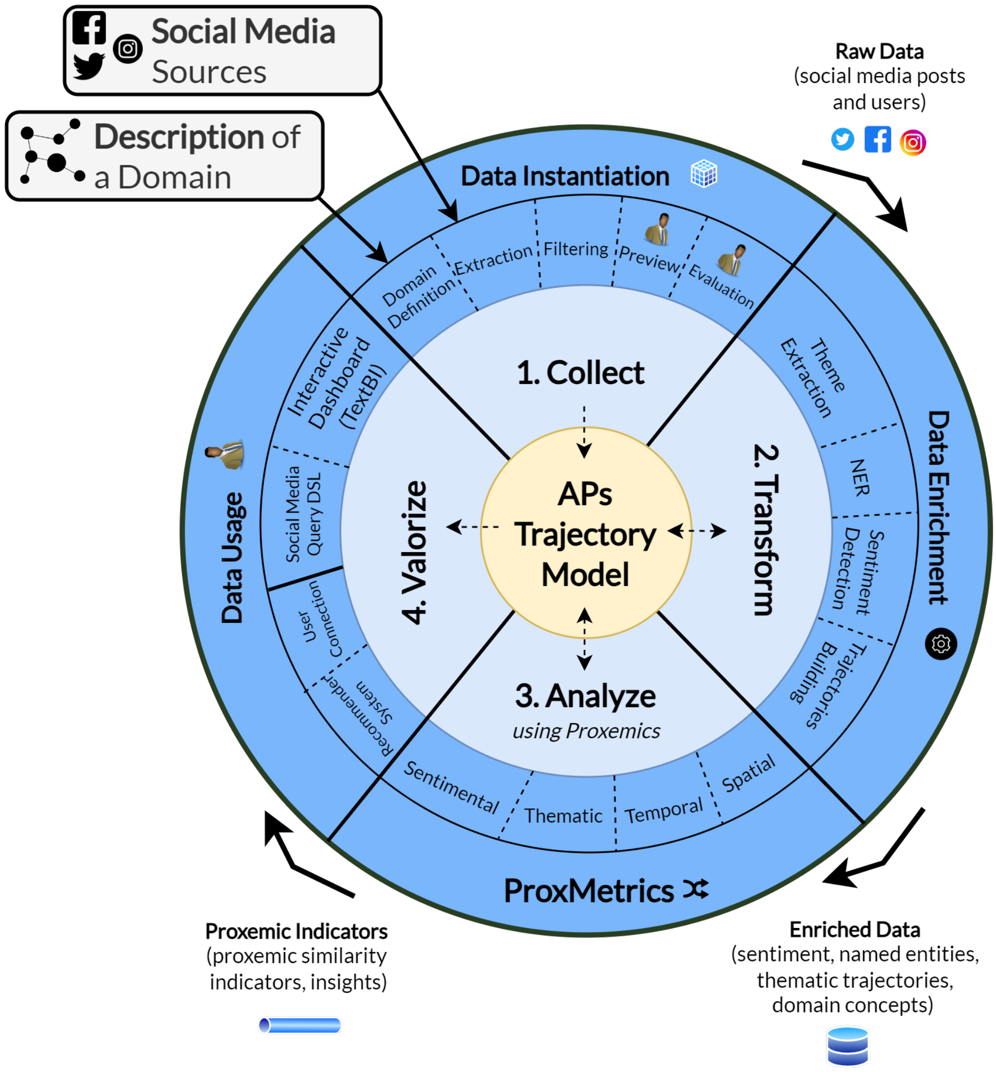


Figure 1.5: Life Cycle of the APs Framework

The end user is involved only in the first phase to define his requirements and evaluate the collected dataset, and during the last one to visualize the results. The work carried out during this thesis enables semi-automatic use of this framework, meaning that actions by computer scientists are still required to transition between phases. However, the future goal is to allow for an automatic framework. Let's now examine the ongoing research challenges that this framework attempts to tackle.

1.4.3 Main Research Challenges and Working Hypotheses

Our goal is to design a framework for processing and analyzing social media data to generate meaningful indicators and insights accessible to non-computer scientists, in a manner that is generic and adaptable across various domains and social media platforms. The research challenges we have identified are linked with the difficulties discussed in Subsection 1.2.2 and include:

- *Challenge 1: Constructing Accurate and Representative Datasets.* Creating datasets from social media that meaningfully analyze real behaviors and phenomena, despite the massive and noisy volume of data on these platforms (refer to Subsection 1.2.2), is complicated. Additionally, the methodology used to build datasets must be applicable across a wide range of domains, spatial areas, and periods, aiming for a generic applicability.
  - *Hypothesis 1*: The use of an iterative, semi-automatic approach that incorporates human feedback at different stages and combines various existing filtering techniques (both content-based and metadata-based) along a semantic domain description could provide a generic and reusable process to build thematic datasets from social media.
- *Challenge 2: Transforming Unstructured Data into Structured Knowledge.* Dealing with the complexities of social media posts, including short, informal messages, poor grammar and punctuation, and the presence of emojis, hashtags, and URLs (refer to Subsection 1.2.2), presents a significant challenge.
  - *Hypothesis 2*: For each specific domain of application, a comparative analysis among the existing NLP techniques (both rule-based and deep learning-based) and language models would allow for the identification of the most suitable ones for this domain.
- *Challenge 3: Performing Well When Annotated Data is Scarce or Inexistent in Multilingual Contexts.* Social media data is inherently multilingual (refer to Subsection 1.2.2). The challenge here is for deep learning-based knowledge extraction modules to perform well when manually annotated training data is scarce or inexistent.
  - *Hypothesis 3*: In addition to the previous hypothesis, leveraging a multilingual training dataset (existing or newly constructed) tailored to the relevant domain can establish a solid foundation for processing multilingual social media data using deep learning-based techniques in few-shot contexts.
- *Challenge 4: Designing Meaningful Domain-Adaptable Indicators for Actionable Insights.* Social media are very diverse and cover various domains (refer to Subsection 1.2.2). Therefore, the indicators we aim to design must be domain-independent and adaptable, to provide valuable insights to stakeholders across various domains.

  - *Hypothesis 4*: Adapting the *proxemics* theory (Hall, 1966; Hall et al., 1968) and proxemic dimensions (Greenberg et al., 2011) to social media could provide a generic and versatile way to model interactions and to produce relevant domain-adaptable indicators.

- *Challenge 5: Presenting Social Media Analyses to Non-Computer Scientists in a Domain-Adaptable Manner.* Making the analyses accessible and understandable to end users without a background in computer science, while being domain-independent, presents another key challenge.

  - *Hypothesis 5*: An interactive dashboard based on four broad dimensions: spatial, temporal, thematic, and personal, correlated with enrichment data such as sentiment and engagement, inspired by design elements from Business Intelligence (BI) platforms, Geographic Information Systems (GIS), and Linguistic Information Visualizations, could provide an accessible (e.g., easy to manipulate) and versatile (e.g., able to satisfy a wide array of domain-specific requirements, such as those in Table 1.1) way for non-computer scientist users to analyze social media data and social media-based indicators in various domains of application.

Addressing these research challenges is critical to the success of the APs Framework. This leads us to the main scientific contributions of this thesis.

1.4.4 Contributions

The main contributions identified within the different phases of the APs Framework, are described as follows. We present four core contributions, some of which are further divided into distinct facets. Each contribution is associated with a specific phase of the APs Framework (refer to Figure 1.5).

Contribution 1: Generic and Iterative Methodology for Constructing Thematic Datasets from Social Media

The initial contribution of this thesis is situated within the *Collect* phase of the APs Framework (Figure 1.5). We propose a generic methodology for building thematic datasets from social media. Many research works gather data from social media, but the extraction processes are often ad-hoc and lack a formal or standardized method. This contribution aims to extend the processes currently used by designing an iterative, generic, and domain-independent approach for building thematic datasets from social media, with three modulable dimensions at its core: spatial, temporal, and thematic. This contribution is linked to the first identified research challenge (Challenge 1).

Contribution 2.1: Novel Multilingual Dataset of Touristic Tweets Covering the *Basque Country* Region, Annotated for Three Knowledge Extraction Tasks

The second contribution falls within the *Transform* phase of our framework (Figure 1.5) and is divided into two parts, both of them linked to Challenge 2 and Challenge 3. The first part is a novel annotated dataset of tweets from the *Basque Country* region. This dataset is multilingual, encompassing French, Spanish, and English, and has been manually annotated for three common knowledge extraction tasks: Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction. Each post in the dataset is annotated with sentiment at the text level, and with locations and fine-grained touristic thematic concepts at the token level, following the definitions set by the *Thesaurus on Tourism and Leisure Activities* of the World Tourism Organization (World Tourism Organization, 2002). This dataset marks a significant milestone as the first of its kind within the tourism domain. While there are existing resources for Sentiment Analysis and NER, they are often too broad and lack the necessary context specific to this domain. Furthermore, there has been no existing dataset dedicated to the extraction of fine-grained touristic thematic concepts. This dataset is instrumental in training deep learning-based models and classifiers to automate NLP processes within this domain.

Contribution 2.2: Comparative Study on Optimal Strategies for Multilingual Analysis of Social Media Content in the Tourism Domain

The second part of Contribution 2 is also linked to the *Transform* phase (Figure 1.5). It unfolds into a dual study. The first entails a comparative study focusing on various NLP techniques (including both rule-based and deep learning-based ones). This study aims to identify the most effective techniques and language models for extracting knowledge from social media data within the tourism domain, specifically for the tasks of Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction outlined previously. The second part of the study focuses on deep learning-based techniques and explores the efficiency of annotated training examples by experimenting with various numbers of examples and dataset sampling methods. This exploration seeks to ascertain the minimal number of annotated training examples required to achieve competitive results across the tasks, employing various training strategies and language models. Given the time-consuming and costly nature of manual data annotation, it is crucial to minimize manual annotation without compromising on quality. Thus, this study aims to identify the optimal balance, beyond which additional data annotation does not significantly enhance results.

Contribution 3.1: Formal Redefinition of the *Proxemics* Theory for Social Media

The third contribution encompasses three distinct parts and is included in the *Analyze* phase of the framework (Figure 1.5). It attempts to address Challenge 4. Initially, we propose a redefinition of the *proxemics* theory (Hall, 1966; Hall et al., 1968; Greenberg et al., 2011) (see Subsection 1.4.1) for application within digital social media spaces (e.g., virtual platforms and environments where individuals and communities interact, share content, and communicate through the internet). We hypothesize that redefining and extending *proxemics*, which is traditionally applied to physical spaces, and its dimensions to characterize social media interactions could provide an effective tool for domain-independent analyses. Therefore, we formally redefine physical proxemic dimensions for their application within digital social media spaces. This novel redefinition of *proxemics* allows us to characterize entities (e.g., users, groups, places, periods, themes) and interactions on social media in a domain-independent manner.

Contribution 3.2: Proxemic-based Trajectory Data Model for Social Media

The second part of Contribution 3 is shared by all steps of the framework (Figure 1.5), although it is mostly associated with the *Analyze* phase. Building upon our formal redefinition of *proxemics*, we introduce a proxemic-based data model along with OCL (*Object Constraint Language*) constraints specifically designed for modeling social media entities, along with their trajectories and interactions, in a domain-independent manner. This model, called the APs Trajectory Model, is multidimensional, drawing upon the five dimensions of *proxemics* (Distance, Identity, Location, Movement and Orientation, DILMO) (Greenberg et al., 2011) redefined for use within social media. It is designed to be source-independent, modular and extensible, to accommodate a wide range of use cases and requirements. Unlike existing social media models, it also incorporates the concept of proximity into digital social media spaces.

Contribution 3.3: *ProxMetrics*, Toolkit for Generating Domain-Adaptable Indicators from Social Media

Based on our formal redefinition of *proxemics* and the proxemic data model, we introduce a modular and generic toolkit accompanied by formulas designed to compute domain-adaptable indicators for social media data analysis. This toolkit, named *ProxMetrics*, enables the expression of these indicators as proxemic similarity metrics between multidimensional social media entities, including but not limited to users, groups, places, themes, and temporal periods. These indicators are customizable, allowing for the modulation of the five dimensions of *proxemics* to address various domain requirements.

Contribution 4: *TextBI*, A Generic and Interactive Dashboard for Visualizing Multidimensional Analyses in Social Media Data

The last contribution is part of the *Valorize* phase (Figure 1.5) and deals with Challenge 5. Here, we introduce *TextBI*, an interactive, generic dashboard designed to present multidimensional analyses (both annotations and indicators) on multilingual social media data. This tool focuses on four core dimensions: spatial, temporal, thematic, and personal, and also supports additional enrichment data such as sentiment and engagement. Multiple views are offered, including frequency, movement, association, and *proxemics*. This dashboard addresses the challenge of facilitating the interpretation of multidimensional social media analyses and indicators by visualizing them in a user-friendly, interactive interface for non-computer scientist users (such as stakeholders in various domains). This tool blends features from various families of visualization tools including Business Intelligence platforms, Geographical Information Systems, and Linguistic Information Visualizations to address their respective limitations in the context of our framework.

1.4.5 Research Fields

This thesis contributes to four distinct research fields, each encompassing specific aspects of the work presented. We have used the ACM Computing Classification System[8](CCS) to classify them.

1. *Web and Social Media Search*: Contribution 1.

[8] https://dl.acm.org/ccs

2. *Natural Language Processing (NLP)*, *Information Extraction (IE)* and *Multilingual Text Mining*: Contributions 2.x.

3. *Decision Support Systems* and *Data Analytics*: Contribution 3.x.

4. *Visualization*, *Human Computer Interactions* and *Interactive Systems and Tools*: Contribution 4.

1.5 Publications

The contributions have been published at the following peer-reviewed international conferences and journals[9]:

P1 M. Masson, C. Sallaberry, R. Agerri, M. N. Bessagnet, P. Roose, A. Le Parc Lacayrelle. (2022). A Domain-independent Method for Thematic Dataset Building from Social Media: The Case of Tourism on Twitter. In *International Conference on Web Information Systems Engineering (WISE 2022)* (pp. 11-20) (Biarritz, France). Cham: Springer International Publishing (CORE Rank: B, ERA Rank: A). [Contribution 1]

P2 M. Masson, P. Roose, C. Sallaberry, R. Agerri, M. N. Bessagnet, A. Le Parc Lacayrelle. (2023). APs: A Proxemic Framework for Social Media Interactions Modeling and Analysis. In *International Symposium on Intelligent Data Analysis (IDA 2023)* (pp. 287-299) (Louvain-La-Neuve, Belgium). Cham: Springer Nature Switzerland (CORE Rank: B, ERA Rank: A). [Contribution 3]

P3 M. Masson, P. Roose, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, R. Agerri. (2024). ProxMetrics: Modular Proxemic Similarity Toolkit to Generate Domain-Adaptable Indicators from Social Media. In *Social Network Analysis And Mining*, 14, 124. Springer. (Impact Factor: 2.8). [Contribution 3]

P4 M. Masson, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose, R. Agerri. (2024). TextBI: An Interactive Dashboard for Visualizing Multidimensional NLP Annotations in Social Media Data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2024)* (pp. 1-9) (St. Julians, Malta). Association for Computational Linguistics. (CORE Rank: A, ERA Rank: A). [Contribution 4]

They have also been published at the following peer-reviewed national conferences and journals:

P5 M. Masson. (2022). Augmented Proxemic Services for Cultural Heritage and Tourism Practices. In *Young Researchers' Forum at INFORSID 2022 (Dijon, France)*. [Presentation of the APs Framework]

P6 M. Masson, R. Agerri, C. Sallaberry, M. N. Bessagnet, P. Roose, A. Le Parc Lacayrelle. (2024). Optimal Strategies for the Multidimensional Analysis of Multilingual Content from Social Media. In *Proceedings of the 42nd Conference on Computer Science for Organizations and Information and Decision Systems (INFORSID 2024)* (Nancy, France). [Contribution 2]

[9] Papers currently in peer review or not directly linked to the subject of this thesis are not listed here.

P7 M. Masson, S. Abdelhedi, C. Sallaberry, R. Agerri, M. N. Bessagnet, A. Le Parc-Lacayrelle, P. Roose. (2023). Interactive Visualization of Touristic Activity Trajectories: Application to Data Extracted from Twitter. In *Workshop "Exploring traces in an all-digital world: challenges and perspectives" at INFORSID 2023 (La Rochelle, France)*. [Contribution 4]

P8 M. Masson, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose, R. Agerri. (2024). TextBI: Interactive Visualization of Multidimensional Data from Social Media. In *Mappemonde. Quarterly Journal on the Geographic Image and the Forms of the Territory*. [Contribution 4]

(A1) Lastly, the *TextBI* platform (Contribution 4) won the 1st prize [10] at the GeoData Challenge during the National Geonumeric Days (GeoDataDays), held at the Reims Convention Center on September 12-13 2023. The event was organized by the *French Association for Geographic Information* (Afigéo) and *DécryptaGéo*.

1.6 Thesis Organization

For the organization of this thesis, we have chosen to adopt the same organizational flow as the APs Framework (see Figure 1.6), meaning that we distribute the contributions across chapters corresponding to the different phases of the framework (which, as a reminder, are *Collect*, *Transform*, *Analyze*, and *Valorize*). Since each stage of the framework deals with quite different domains of computer science, it seemed logical to us to group related elements. We believe that this organization allows readers to better understand our contributions in the context of the framework.

We illustrate our proposals with examples in the tourism domain. The genericity of the contributions, adapted to any domain will be demonstrated at the end, in a dedicated chapter (Chapter 6). Thus, the thesis is organized as follows.

Chapter 2 deals with the initial data collection step (*Collect*). We start by reviewing existing techniques for extracting data from social media. Many existing works rely on social media data, yet the current extraction techniques often lack a formalized, standardized process and are mostly ad-hoc. This chapter, therefore, proposes a novel collection method, designed to build thematic datasets from social media (Contribution 1). It leverages three dimensions: spatial, temporal, and thematic. Its main originality is that it is generic, meaning that it can adapt to any domain of interest and follows an iterative flow, incorporating human feedback at different steps of the process. Our approach is experimented with and evaluated through the development of a multilingual dataset focused on tourism within the touristic region of the *Basque Country*, leveraging X/Twitter as the data source. This dataset undergoes both quantitative and qualitative evaluations to demonstrate the method's effectiveness and its ability to produce valuable thematic datasets.

Chapter 3 focuses on the transformation of previously collected data (*Transform*) from unstructured raw text to structured annotations. We review existing NLP techniques for three knowledge extraction tasks: Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction with a focus on deep learning-based methods. Additionally, we review existing training corpora for these tasks. Due to a lack of existing multilingual training resources in the domain of tourism, we propose and describe the process of creating a novel dataset, manually annotated at the text level with sentiments and at the token

[10] https://www.geodatadays.fr/page/GeoDataDays-2023-Les-Challenges-Geodata/139

levels with locations and thematic touristic concepts (Contribution 2.1). This dataset serves as a basis for a comparative study between different training techniques and language models to determine what are the best strategies for these three knowledge extraction tasks in the domain of tourism. Additionally, for each deep learning-based technique, we try to determine the tipping point at which annotating more data does not yield significantly better results (Contribution 2.2). The objective is for researchers to not manually annotate too much training data, a time-consuming and costly process.

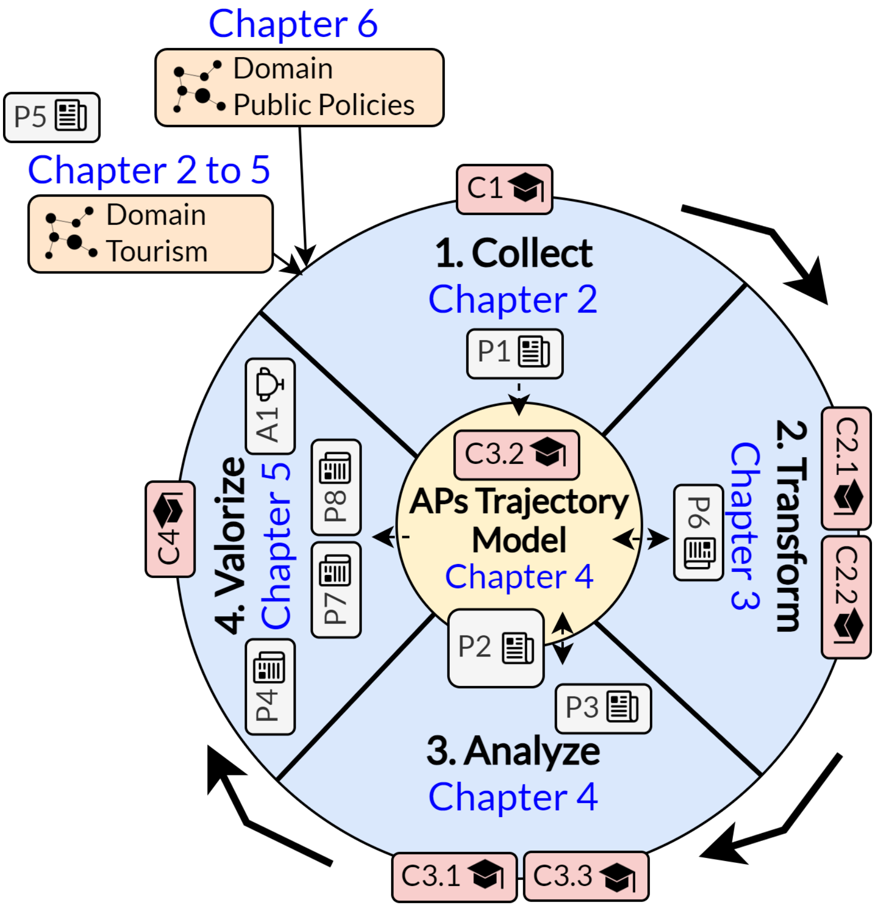


Figure 1.6: Overview of Chapters, Contributions, and Publications Mapped to the APs Framework (*refer to Subsection 1.4.4 and Section 1.5 for publications and contributions identifiers*)

Chapter 4 represents the subsequent phase (*Analyze*). This chapter aims to transform the previously annotated dataset into actionable knowledge and indicators useful for domain stakeholders. To achieve this, we hypothesized that applying the *proxemics* theory (Hall et al., 1968), could be beneficial. Thus, we begin by reviewing the existing applications of *proxemics* across various fields and propose a novel, digital redefinition of *proxemics* (Contribution 3.1) to characterize social media entities and interactions. This redefinition leads to the creation of a new proxemic model: the APs Trajectory Model (Contribution 3.2). This model is generic, and extensible, and serves as the modular data backbone of our framework. Moving forward, we introduce a toolkit, *ProxMetrics*, designed to calculate domain-adaptable indicators based on this model (Contribution 3.3). These indicators, expressed as proxemic similarity measures, are developed following a thorough review of existing similarity measurement techniques. Experiments were carried out in the tourism domain using the previously collected and annotated dataset, as well as the requirements gathered from local tourism offices.

Chapter 5 discusses the final phase of our framework (*Valorize*). We review existing visualization tools for decision support, including Geographic Information Systems (GIS), Business Intelligence (BI), and Linguistic Information Visualizations, used to present data analyses to non-computer scientist end users, such as domain stakeholders. These tools can assist them in the decision-making process by providing informed insights. From this comprehensive review, we notice that all of these tools have limitations that do not allow them to address our framework's requirements and the specificities of social media data. Therefore, we introduce an interactive, generic dashboard: *TextBI* (Contribution 4). It combines elements from the aforementioned technologies: it incorporates advanced spatial views from GIS, leverages the advanced interactivity and combined filtering capabilities of BI tools, and integrates text-based aspects from Linguistic Information Visualizations into an accessible, multidimensional dashboard. Moreover, *TextBI* incorporates the previously calculated proxemic similarity indicators and presents them in a user-friendly manner to end users. The platform underwent qualitative evaluation, notably through a collaboration with local tourism offices.

Chapter 6 aims to demonstrate the genericity of the framework and our proposals on another data source and domain of application. We diverge from focusing solely on X/Twitter and the tourism domain to experiment with our processes within a new domain of application: local public policies. This additional experiment leverages data from municipality review platforms (such as *bien-dans-ma-ville.fr*[11]) about municipalities all around *France*. We validate that each step of the framework adapts to this new data source and domain of application.

Chapter 7 concludes this thesis and presents various future research avenues that we plan to explore, as well as possible extensions of the APs Project, notably through a French-Japanese postdoctoral project.

[11] https://www.bien-dans-ma-ville.fr

Chapter 2

Collect

Toward a Generic and Iterative Methodology for Constructing Thematic Datasets from Social Media

"Data! Data! Data! I can't make bricks without clay!"
— Sir Arthur Conan Doyle, Scottish Writer and Doctor

In the rapidly evolving digital age, the proliferation of social media platforms has started a new era of data abundance. This unprecedented availability of data presents both a remarkable opportunity and a serious challenge for researchers. Indeed, while the volume of data is important, acquiring accurate and relevant data is even more crucial for conducting meaningful analyses. This chapter, associated with the *Collect* phase of the APs Framework (see Figure 1.5), addresses the challenge of constructing accurate and exhaustive thematic datasets from social media.

Raw Dataset
Social Media Posts and Users
Collect
Transform
Analyze
Valorize

We begin with a brief introduction (Section 2.1) followed by a review of existing works that required the construction of social media datasets and the collection techniques they employed, categorized into three types: spatial, temporal, and thematic, noting a lack of standardized collection methods (Section 2.2). The methods used are mostly ad-hoc and tailored to requirements specific to each project. From this observation, and to address the requirements of our framework, we propose a novel generic and iterative methodology for constructing thematic datasets from social media (Section 2.3, Contribution 1). This contribution is positioned in the *Web and Social Media Search* research field. We hypothesize that the use of a semi-automatic approach, incorporating human feedback at various stages and combining existing metadata-based and content-based filtering techniques with a semantic domain description, could provide a generic and reusable process for building thematic datasets from social media. The methodology is then experimented through the development of a dataset focused on tourism within the touristic region of the *Basque Country*, leveraging X/Twitter[1] as the data source (Section 2.4). Finally, we evaluate it using both qualitative and quantitative metrics (Section 2.5) and propose perspectives for improvement (Section 2.6). The methodology introduced in this chapter has been published at the following international conference:

- M. Masson, C. Sallaberry, R. Agerri, M. N. Bessagnet, P. Roose, A. Le Parc Lacayrelle. (2022). A Domain-independent Method for Thematic Dataset Building from Social Media: The Case of Tourism on Twitter. In *International Conference on Web Information Systems Engineering (WISE 2022)* (pp. 11-20) (Biarritz, France). Cham: Springer International Publishing (CORE Rank: B, ERA Rank: A).

2.1 Introduction: Beyond Volume, The Importance of Accurate Datasets for Social Media Analytics

In an era where social media platforms generate an overwhelming flood of data every second, the capability to capture, analyze, and extract meaningful insights from this data has become a pivotal challenge in research, marketing, policy making, and beyond (Kaplan and Haenlein, 2010). While the vastness of these datasets is undeniable, offering a seemingly endless reservoir of information, the true value of social media analytics lies not just in the quantity of data but, crucially, in its quality.

A primary challenge in handling social media data, characterized by its user-generated and uncurated nature, is its massive volume accompanied by high levels of noise (El Abaddi et al., 2011). At the end of 2023, more than 500 million tweets were sent daily (Statista, 2024) and a study by the *Pew Research Center* estimates that around 36% of online adults aged between 18 and 29, and 23% of online adults aged between 30 and 49 use X/Twitter in the US (Shannon Greenwood and Duggan, 2016). This volume, coupled with the presence of noise and irrelevant information such as bots, duplicate content, and off-topic discussions, can significantly bias analyses, leading researchers to potentially draw incorrect conclusions, thereby forming decisions and strategies on a flawed foundation. For instance, it is estimated that over 50% of tweets consist of "*pointless babble and irritating spam*" (Liu et al., 2016). Hence, accurate datasets are fundamental to high-quality social media analytics, ensuring that analyses accurately reflect genuine user behaviors and sentiments, free from the distortions of irrelevant or misleading data. In domain-specific research, where the focus is on extracting data related to specific themes (for example, *tourism*), the precision of datasets becomes even more critical.

Building accurate datasets can be a complex task. Various social media platforms introduce different methods for extracting data, often through rate-limited *Application Programming Interfaces* (APIs) with limited querying capacities. Consequently, each request must be carefully selected to avoid deadlock. Additionally, as part of the APs framework, there is a requirement to be generic in application domains, enabling the construction of datasets spanning a wide and varied range of domains.

In summary, we are confronted with two key challenges: (1) the ability to filter the massive

[1] https://twitter.com

and noisy data on social media to extract relevant subsets that will allow us to analyze behaviors and phenomena reflective of the real world, and (2) the capacity to do this across any domain of application and social media platform (*genericity*). Let's now examine the approaches and techniques that existing research works gathering social media data use to build their datasets.

2.2 Related Work: Existing Dataset Building Approaches from Social Media

We will now review the most common approaches used to build datasets from social media. Table 2.1 presents a collection of research works gathering social media data. We noticed that the extraction criteria they use are often applied to the same three core dimensions (refer to Figure 2.1): spatial (Subsection 2.2.1), temporal (Subsection 2.2.2), and thematic (Subsection 2.2.3). These dimensions are used to selectively extract posts matching a given spatial footprint, targeted temporality, or precise semantics. They apply to both the content of the posts and the metadata associated with them.

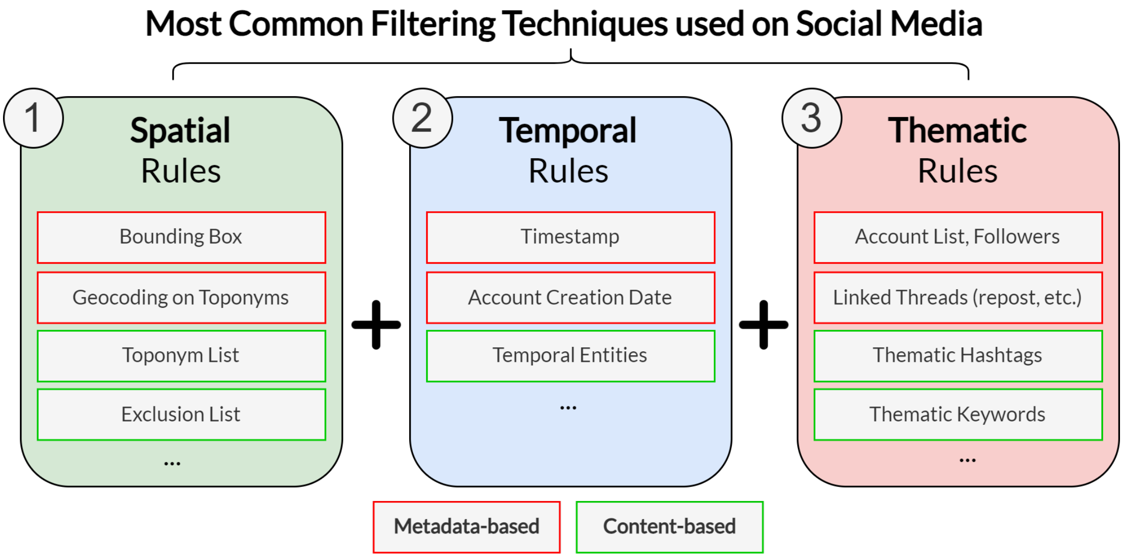


Figure 2.1: Most Common Filtering Techniques Used to Extract Data from Social Media

2.2.1 Spatial Filtering

For *spatial filtering* (Figure 2.1, (1)), two different approaches are generally used and sometimes combined to obtain better results. First, filtering can be performed on the spatial metadata attached to the post, either by providing the social media API with a bounding box encompassing the study area (Paolanti et al., 2021), or by basing it on a precise location (*referenced with a latitude/longitude*) with a radius around it (this is called *location nearness*) (Scholz and Jeznik, 2020; Zotova et al., 2021). The main advantage of this metadata-based approach is that it is extremely accurate, as the location is generally determined using the *Global Positioning System* (GPS) of the device. The biggest problem that arises is the incredibly low number of users who enable geotagging and therefore the limited number of posts that have metadata attached. For X/Twitter, it was estimated that around 1% to 2% of tweets have spatial metadata attached to them (Sloan and Morgan, 2015).

Thus, relying solely on metadata misses many potentially relevant posts whose authors have not enabled geotagging.

The second approach used is applied to the content of the post and is done via toponym filtering (Viñán-Ludeña and de Campos, 2021; Shimada et al., 2011; Scholz and Jeznik, 2020). A list of place names, such as municipality names or points of interest (POIs), along with their associated abbreviations and translations, is compiled, and all posts containing them are extracted. The main drawback of this methodology is the potentially large amount of noise (e.g., post containing a toponym but not related to the study area, or too common toponyms). For example, if a study is focused on *Paris, Texas*, posts mentioning *Paris* might refer to *Paris, France*, leading to irrelevant data being included. Several approaches are used to mitigate this problem, for example by combining this list of toponyms with exclusion lists (Viñán-Ludeña and de Campos, 2021) (e.g., keywords that should not appear in a post containing a toponym), this is particularly useful when several places with the same designation exist to disambiguate them.

2.2.2 Temporal Filtering

For the *temporal filtering* (Figure 2.1, ②), the extraction is almost always performed based on the timestamp metadata, as seen in Table 2.1. This methodology is straightforward to implement and is usually quite accurate. Exchanges on social media often occur in real time, making it unnecessary to establish a complex system for extracting temporal entities. Additionally, almost all social media platforms provide such timestamp metadata.

However, for highly fine-grained temporal tasks, it may be necessary to infer temporal relationships in posts, such as *yesterday*, *tomorrow*, *three days ago*, etc., using a system like the ones proposed in Alfattni et al. (2020) or as part of the *TempEval* challenge (Verhagen et al., 2009).

2.2.3 Thematic Filtering

When it comes to the *thematic filtering* (Figure 2.1, ③), several approaches are used. First, content-based approaches use thematic keywords directly related to the domain of study (Grant-Muller et al., 2015; Agerri et al., 2021; Zubiaga et al., 2016; Basile et al., 2018), for example, local event names (Shimada et al., 2011). The use of too precise keywords can restrict too much the number of returned posts, so some research works associate several words together to establish multi-criteria filtering rules. More and more social media (such as *X/Twitter*, *Instagram*, *Facebook* among others) have the concept of "*hashtag*", keywords preceded by the # character allowing to identify topics of discussion. It is a filtering tool that is often used for extraction (Cignarella et al., 2020; Zotova et al., 2021; Hürlimann et al., 2016). A set of hashtags related to the theme is defined and all the posts containing them are extracted. More advanced syntactic rules can also be set up depending on the requirements (e.g., extracting posts containing at least one vowel, written in the Cyrillic alphabet, etc.), but may need to be applied locally to an already extracted set of posts due to the lack of support in the social media filtering API.

Other thematic filtering methods are applied to the metadata and include but are not limited to: the language of the post and its source (e.g., content sent from a mobile or fixed device, shared via another social media, etc.).

Reference	Source	Objective	Spatial Criteria		Temporal Criteria	Thematic Criteria	
			Metadata-based	Content-based	Metadata-based	Metadata-based	Content-based
Paolanti et al. (2021)	X/Twitter	*Tourism in Cilento, Italy*	• Bounding Box		• Timestamp		
Viñán-Ludeña and de Campos (2021)	X/Twitter	*Tourism in the province of Granada, Spain*		• Place names • Place exclusion	• Timestamp		• High-frequency hashtags • Thematic hashtags • Thematic keywords
Shimada et al. (2011)	X/Twitter	*Tourism information in a local Japanese municipality*		• Place names • Place abbreviations	• Timestamp		• Event names
Scholz and Jeznik (2020)	X/Twitter	*Visitor flows in Styria, Austria*	• Location nearness on geocoded names	• Municipality names	• Timestamp		• Tourism keywords
Basile et al. (2018)	X/Twitter	*Multiple*			• Timestamp	• Language filter	• Thematic keywords • Thematic hashtags • Syntactic rules
Chiruzzo et al. (2020)	X/Twitter	*Spanish Humor*			• Timestamp	• From a list of humor-related accounts	• Humor-related hashtags
Hürlimann et al. (2016)	X/Twitter	*Sentiment analysis for the Brexit referendum*			• Timestamp	• Brexit-related accounts	• Brexit hashtags • Common Brexit keywords
Zotova et al. (2021)	X/Twitter	*Stance detection*	• Location nearness		• Timestamp	• Most active linked accounts • Language filter	• Thematic hashtags • Thematic keywords
Cignarella et al. (2020)	X/Twitter	*Stance detection*			• Timestamp	• Source-based filtering • No media • Maximum one post per account	• Thematic hashtags • Thematic keywords • Associated replies, quotes, and retweets
Van Bruwaene et al. (2020)	X/Twitter	*Cyberbullying in social media*					• Cyberbullying-related hashtags
Zubiaga et al. (2016)	X/Twitter	*Rumors spreading on social media*				• Associated threads	• Hot topics keywords
Agerri et al. (2021)	X/Twitter	*Vaccine skepticism*				• Most active Basque account	• Vaccine hashtags • Vaccine keywords
Lewis et al. (2008)	Facebook	*University students' use of social media*				• Profile privacy settings • Account name	
Santia and Williams (2018)	Facebook	*Veracity assessment of news and social bot detection*				• News outlets' accounts • Associated comments and reactions	
Zarei et al. (2020)	Instagram	*Covid-19*			• Timestamp	• Associated comments and reactions	• Covid-19 hashtags
Turcan and Mckeown (2019)	Reddit	*Identification of stress*			• Timestamp	• List of stress-related subreddits	
Rogers et al. (2018)	VKontakte	*Sentiment analysis in Russian*			• Timestamp	• From users part of Maidan-related communities • Post outside these communities • Posts having comments	• Excluding political keywords • Syntactic criteria (e.g., length, contains Cyrillic alphabet characters, less than four hashtags)

Table 2.1: Overview of Common Filtering Approaches Used to Build Datasets from Social Media

To reduce noise, some research works use only posts from a pre-selected list of accounts known to validate certain desired features (for example, known for posting humor like in Chiruzzo et al. (2020), known for speaking regularly about a specific topic, etc.). Sometimes, and depending on the extraction requirements, associated replies and comments can also be extracted (Cignarella et al., 2020).

For each dimension, the granularity of the filtering process is typically determined based on the specific requirements of the project. Projects requiring a vast corpus of posts may exhibit greater tolerance for noise, whereas those involving smaller datasets usually necessitate the processing of posts that are strictly pertinent to the domain of application.

2.2.4 Existing Frameworks for Data Collection

In recent years, efforts have been made to design more generic processing techniques for social media data (Morgan and Van Keulen, 2014; Sathick and Venkat, 2015). However, these are mainly intended for higher-level Information Extraction (IE) and do not emphasize the methodology used for dataset building. They usually consist of NLP processing pipelines applied to an already pre-selected, pre-extracted set of social media posts. They do not provide domain-independent, formalized strategies to collect these posts.

Similarly, we have witnessed the appearance of comprehensive, multi-step frameworks for extraction from social media. However, these are typically focused on a specific domain of application or specific data source. For example, Jayawardhana and Gorsevski (2019) proposed an ontology-based framework for extracting spatio-temporal influenza data using X/Twitter. Parekh et al. (2018) proposed a reusable data collection pipeline for the study of terrorism and jihadists' behaviors on social media based on seed accounts. Hussain et al. (2017) proposed a generic methodology to collect data, but specifically targeted at blog websites. Other works have attempted to use ontologies as a way to model and facilitate goal-oriented queries on social media, such as Izhar et al. (2016). However, these methods are applied to already pre-collected posts too and, therefore, do not propose strategies to collect the posts (e.g., they propose strategies to annotate posts but not to collect them).

Another challenge was highlighted by Campan et al. (2018). They demonstrated that when querying the X/Twitter collection API with concurrent processes to filter posts for highly popular terms, the processes tend to return almost identical posts, despite only receiving a sample of all posts that match the keywords. Conversely, for less popular keywords or hashtags, X/Twitter is more likely to provide a complete set of matching posts. We suppose that this issue might arise on other social media platforms as well; therefore, this underscores the importance of employing a diverse array of domain-relevant keywords or hashtags to ensure a comprehensive collection of posts.

2.2.5 Discussion

As we can see, when it comes to dataset building, each research project usually comes with its own ad-hoc extraction and filtering process. While some common techniques are shared, such as timestamp filtering, usage of keywords and hashtags, etc., there is no formalized and domain-independent approach or procedure that could be used as a basis to facilitate the construction of thematic datasets from social media.

Some studies propose more sophisticated, reusable techniques for dataset construction, yet these typically fall into one of those categories: (1) they are applied to a pre-selected set of posts, thereby not addressing the challenge of extracting posts directly from social media platforms with the associated challenges (*rate limit*, *limited queries*, etc.), or (2) they are focused on specific domains of application or data sources, which makes the approaches proposed difficult to adapt for use in significantly different domains and other social media contexts.

We will now build on these existing works to propose a formalized, generic methodology for constructing thematic datasets from social media.

2.3 A Generic and Iterative Methodology for Constructing Thematic Datasets from Social Media

We propose a formalized methodology for building thematic datasets from social media (Masson et al., 2022). It aims to formalize a generic approach for extracting data from social media sources related to a given domain and, when applicable, a specific period or spatial area. This methodology is designed around the following properties:

- *Multi-dimensional:* This methodology is based on three abstract dimensions: spatial, temporal, and thematic. The presence of all these dimensions is optional; our methodology supports combining two of them or even using only one.
- *Generic:* It has been designed to be implementable with any social media that features a post system and a way to query those posts. This genericity also extends to languages, enabling the methodology to work with multiple languages.
- *Domain-independent:* The main feature of our methodology is its support for any domain for the future dataset (the domain has to be defined through a semantic vocabulary). Current collection techniques are closely linked to their themes; we aim to offer an alternative that is independent of the theme.
- *Iterative and incremental:* Our methodology is designed around an iterative and incremental process, i.e., each iteration aims at refining subsequent iterations to produce a dataset as exhaustive as possible with minimal noise.

It is important to note that our contribution is not an automated software platform dedicated to the collection of social media posts. Instead, it provides an extraction methodology aimed at guiding social media researchers, regardless of their domain, in building datasets. The exact implementation of this methodology is meant to slightly differ depending on the social media used and the specific requirements of each project, thereby guiding the implementation process.

For a social media platform to be compatible with our methodology, it must offer a mechanism for users to retrieve posts, either through API access or web scraping. The broader and more advanced the query functionalities provided, the easier it will be to implement our methodology, as many steps can then be delegated to the social media platform itself. Furthermore, in the case of datasets with a thematic dimension, it must be feasible to describe this dimension using a semantic vocabulary (e.g., through a dictionary, a thesaurus, or an ontology).

We will now explain how the methodology works by focusing on four aspects of it: the dataset definition (Subsection 2.3.1), the filtering of the flow of posts (Subsection 2.3.2), the discovery of new toponyms and vocabulary (Subsection 2.3.3), and lastly the dataset preview and iteration process (Subsection 2.3.4).

2.3.1 Defining the Dataset

The first step is to define the future dataset along these three core dimensions.

1. The *spatial dimension* refers to the geographical scope of the data. Depending on the project's requirements, it may encompass a municipality, a region, a country, linguistic zones, or it may be absent if the collection process is intended to be global.

2. The *temporal dimension* pertains to the time frame of the data to be collected. It can be specified in various ways: as a straightforward temporal interval, a specific date, or even more flexibly as seasons, days of the week, weekends, etc.

3. The *thematic dimension*, often the most crucial, defines the subject of the dataset. It is characterized by vocabulary related to the theme and is used to filter posts that match the studied domain of application. The complexity of this vocabulary can vary depending on the complexity of the domain, ranging from simple lists of keywords to thesaurus or ontologies.

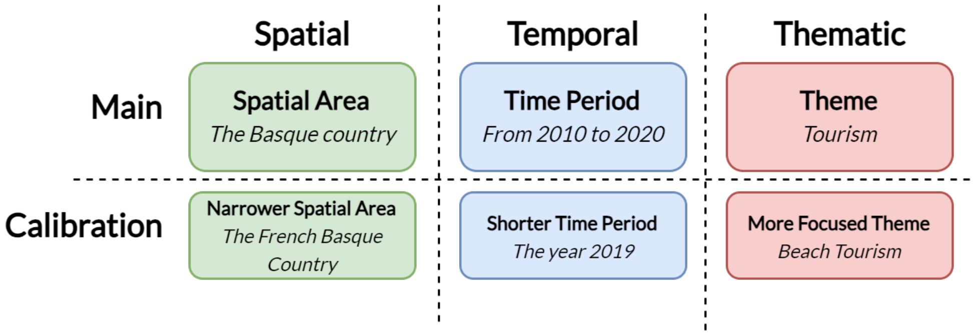
We opted to build our methodology around these dimensions because, after reviewing related works in dataset construction, we concluded they are well-suited for a wide range of use cases (see Section 2.2). Furthermore, most social media platforms offer features that correspond to these dimensions, such as timestamping, geolocation of posts, and textual content for semantic analysis, making the extraction process widely applicable.

It is important to note that not all dimensions have to be present for every project. Thus, datasets could be defined in solely spatial-thematic or temporal-thematic ways, see Table 2.2.

Definition Type	**Resulting Dataset Definition**
Spatial	Posts pertaining to a specific geographical area
Temporal	Posts within a specified timeframe
Thematic	Posts related to a particular theme
Spatio-Temporal	Posts from a geographical area within a timeframe
Spatio-Thematic	Posts from a geographical area related to a theme
Tempo-Thematic	Posts within a timeframe related to a theme
Spatio-Tempo-Thematic	Posts within a timeframe and geographical area related to a theme

Table 2.2: Combination of Dimensions and their Resulting Dataset Definitions

However, it is important to keep in mind that social media often sets limits on the number of posts that can be retrieved during a given period (Anderson et al., 2019). Therefore, it is necessary to define the dataset and set the dimensions taking into account these limits. A dataset definition that is too broad and unrestricted will lead to a very long collection process and a massive resulting dataset. Finding the right balance in the dimensions is sometimes difficult, especially when working on a topic with which one has little or no experience. This is why our methodology is iterative and incremental; users can refine the dimensions later in the process, as much as they want, until they are satisfied with the resulting dataset. There are many ways to evaluate this satisfaction: preview by an expert of a subset of the dataset, implementation of quantitative metrics, etc. We will discuss this further in Section 2.5.



	Spatial	Temporal	Thematic
Main	Spatial Area *The Basque country*	Time Period *From 2010 to 2020*	Theme *Tourism*
Calibration	Narrower Spatial Area *The French Basque Country*	Shorter Time Period *The year 2019*	More Focused Theme *Beach Tourism*

Figure 2.2: Example of a Dataset Definition and its Associated Calibration Dataset

After completing this definition step (Figure 2.2, *Main*), it is recommended, though not mandatory, to define a calibration dataset (Figure 2.2, *Calibration*). This is a subset of the main dataset (e.g., restricted to a narrower region or a shorter period) with a more restrictive definition that will be used to calibrate the main, wider collection process. The goal is to refine and evaluate the implemented process faster on a small dataset and then save time on the main one.

2.3.2 Filtering the Flow of Posts

Once the datasets have been defined, we can now move to the collection process. Figure 2.3 shows an overview of the collection methodology we have designed.

Four successive filtering steps are applied: *Pre-Filtering*, *Temporal Filtering*, *Spatial Filtering* and *Thematic Filtering*. These steps were selected according to Table 2.1 and are applied sequentially to two sets of posts with different features.

1. *Geotagged posts*: These are the posts whose authors have manually activated geolocation. The latter is calculated by the GPS of the device and can therefore be considered reliable. However, these constitute only a fraction of the total number of posts (e.g., about 1% to 2% according to Sloan and Morgan (2015) for X/Twitter). This reduced set of posts is handled first.

2. *Other posts*: This category includes all non-empty posts, except the aforementioned ones.

Each set is extracted following the procedure described in Figure 2.3. The order of the steps (Figure 2.3) is only indicative and may vary depending on the circumstances. It is necessary to consider which steps can be delegated to the internal filtering system of social media and which ones must necessarily be carried out locally (the latter have to be processed last). There is no universal answer to this question; it depends on the extent of the search functionality of the chosen social media and the complexity of the dimensions defined (e.g., having a simple timespan as a temporal dimension will be easier to implement than day-of-the-week-based criteria because social media natively support timespan filtering through their public APIs).

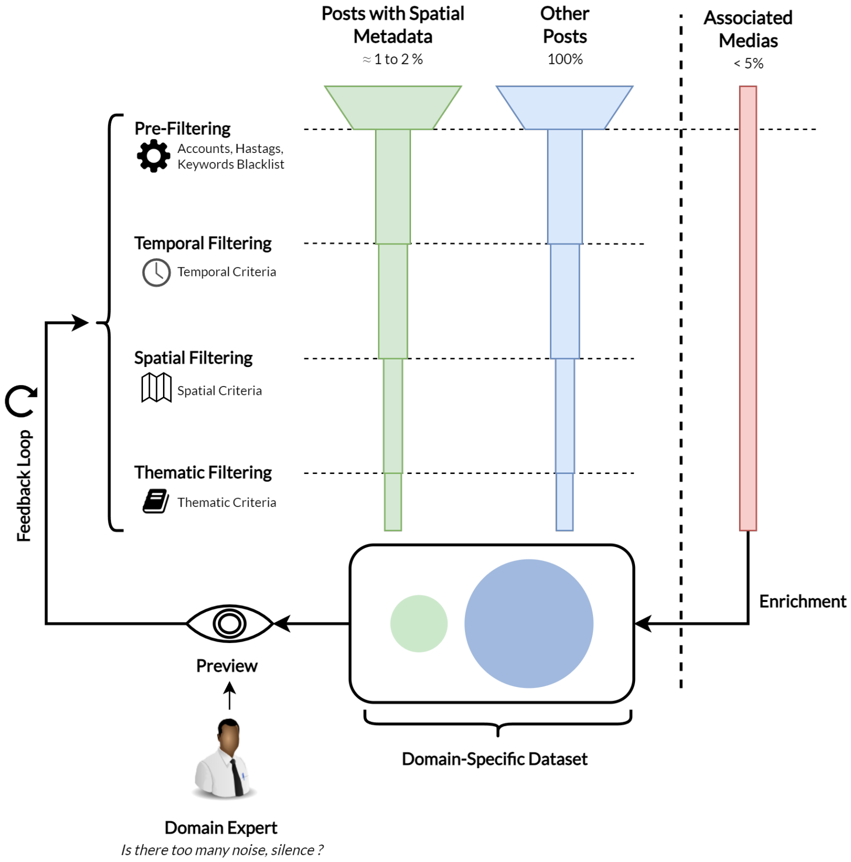


Figure 2.3: Simplified View of the Filtering Process

Pre-Filtering

This step (Figure 2.3, *Pre-Filtering*) aims to exclude accounts, keywords, or hashtags that we know associated posts should not appear in our final dataset while being prone to fit within our dimensions. For example:

- Excluding professional, institutional, or promotional accounts.
- Excluding problematic keywords or hashtags.
- Excluding posts in a certain language.
- Excluding reposts, quotes.
- Excluding posts below a certain length or containing no media.
- Differentiating toponymic homonyms.

For instance, this step is useful to exclude places with the same name but unrelated to each other (*toponymic homonyms* (Caldwell, 2008)), which could distort the spatial filtering process (e.g., two municipalities with the same name in two different areas). Some social media platforms have an exclusion operator in their search system, which should be used for this step. Typically, these criteria are not known at the beginning of the process, so this step is initially empty. It will be filled in during future iterations.

The following steps may be optional depending on the dimensions defined for the dataset. For a more efficient collection process, it is advisable to order them from the dimension believed to be the most restrictive to the least restrictive (to process as few unnecessary posts as possible) and to delegate as much filtering as possible to the social media's API system, thus limiting the number of posts to be filtered out locally.

Temporal Filtering

Temporal filtering (Figure 2.3, *Temporal Filtering*) is conducted using the timestamp of the posts. This technique has been extensively used and covers most use cases. Social media function as instantaneous forums for informal exchange, meaning users generally discuss the present moment, coinciding with their time of posting. The proliferation of mobile media (e.g., *phones, tablets*) has further accentuated this phenomenon, eliminating the requirement to wait to access a computer to publish content. Users can now post anywhere and at any time about a given topic.

In specific cases, a temporal entity detection system could be implemented to analyze the content of the posts, like Alfattni et al. (2020). This would allow for more precision in identifying discussions about past or future events (e.g., *"yesterday," "next week," etc.*). The main drawback of this approach is that social media platforms do not natively support it, meaning it necessitates exporting a very large number of potentially irrelevant posts and then conducting the temporal filtering on them locally.

Spatial Filtering

For spatial filtering (Figure 2.3, *Spatial Filtering*), the approach differs based on whether the posts have spatial metadata. For posts with spatial metadata, which constitute a minority (few users use this feature), we propose filtering directly based on the spatial position of the post using an area polygon or a bounding box. This method, relying on device GPS, is highly accurate and minimizes the risk of noise (e.g., importing posts outside the study area). Furthermore, an increasing number of social media platforms support the use of this feature in their post collection APIs, which facilitates this step.

For the second set of posts, those without spatial metadata (non-geotagged), which represent the overwhelming majority, the approach is somewhat different. We rely on the content of the post (e.g., the text) for selection. A list of toponyms within the study area (e.g., names of places, points of interest, etc.) is used as input and matched within the post content to determine which posts pertain to the area (Shimada et al., 2011). The risk of noise is higher with this method, but our iterative approach allows for further refinement of this step in the process.

Thematic Filtering

The aim of this step (Figure 2.3, *Thematic Filtering*) is to retain only the posts that relate to a specific theme, which is defined using a particular vocabulary. As mentioned previously, the theme can be defined as a simple list of keywords, a dictionary, or even a thesaurus or ontology (Guarino et al., 2009) in more advanced cases. The filtering process involves aligning these semantic representations with the text of the post. Several methods can be employed depending on the requirements and the semantic representation used, such as entity linking (Cossin et al., 2018) or semantic annotations (Uren et al., 2006) (e.g., tagging documents with relevant concepts). Finally, after processing the two sets of posts, it is essential to eliminate potential duplicates.

Enrichment with Associated Media

Lastly, associated media (see Figure 2.3, *Associated Medias*), including pictures, audio, and video attached to posts, can enrich the collected dataset. For instance, computer vision techniques like *Google Cloud Vision* (Mulfari et al., 2016), can extract thematic keywords and locations from images. These can be compared to the list of toponyms and semantic data, providing an additional variable for filtering (in addition to posts' content and metadata).

2.3.3 Discovering New Vocabularies and Toponyms

One of the original aspects of our methodology is the proposal of an architecture that enables the discovery of new vocabularies and toponyms, as well as the evaluation of currently used ones to incrementally enhance the filtering process. Figure 2.4 illustrates a more detailed view of the collection process. The discovery of new vocabulary occurs at two important points in the process:

1. Using the outcomes of the spatial and thematic filtering applied to the first set of posts (*geotagged posts*) to refine the filtering for other posts (Ⓐ and Ⓑ in Figure 2.4).

2. Between each global iteration (*Feedback Loop* in Figure 2.4).

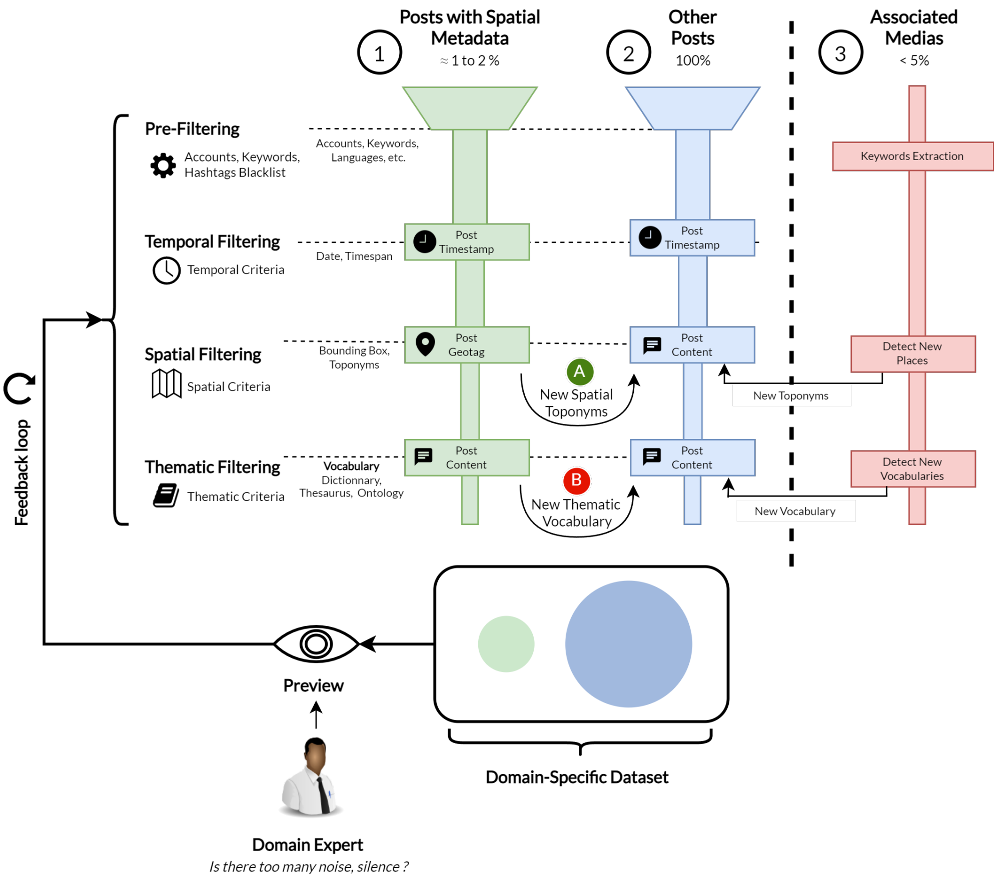


Figure 2.4: Detailed View of the Filtering Process

Regarding toponyms (refer to Figure 2.5 and Figure 2.4, *Thematic Filtering*), we identify and store all toponyms found within the posts immediately following the spatial filtering step (Figure 2.5, ①). After processing the posts, we check if they are already included in our toponym list and, if not, we can proceed to geocode them automatically to determine if they are contained within the study area (Figure 2.5, ②). If certain toponyms are frequently used, related to the study area but not present in the toponym list, we can proceed to automatically add them to it for the next iteration (Figure 2.5, ③). Conversely, some toponyms may be excluded if they prove irrelevant (e.g., no occurrence in the extracted posts), to save processing time and reduce the number of queries sent to the social media API.

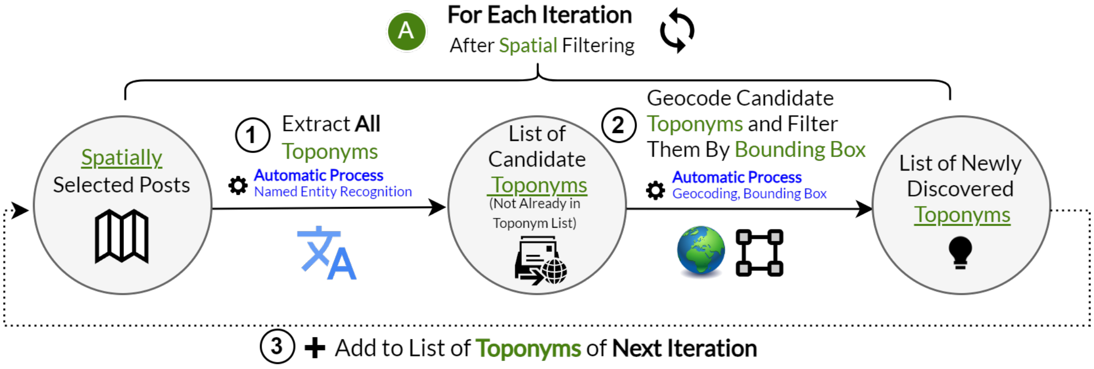

A
For Each Iteration
After Spatial Filtering
Spatially Selected Posts
1 Extract All Toponyms
Automatic Process
Named Entity Recognition
List of Candidate Toponyms
(Not Already in Toponym List)
2 Geocode Candidate Toponyms and Filter Them By Bounding Box
Automatic Process
Geocoding, Bounding Box
List of Newly Discovered Toponyms
3 + Add to List of Toponyms of Next Iteration

Figure 2.5: The Toponym Discovery Process

A similar approach can be adopted for the thematic dimension (see Figure 2.6 and Figure 2.4, *Thematic Filtering*), but it requires manual, human review. Instead of identifying toponyms, this involves performing NLP tasks such as lemmatization and PoS (*Part of Speech*) tagging on the thematically filtered posts to extract frequent terms (Figure 2.6, ①). Then, a domain specialist can analyze frequent nouns, verbs, and adjectives to discover new thematically related vocabulary that is not already contained in the semantic resource (Figure 2.6, ②). These new terms can then be integrated into the current thematic vocabulary (Figure 2.6, ③). This process is performed manually by a domain specialist.

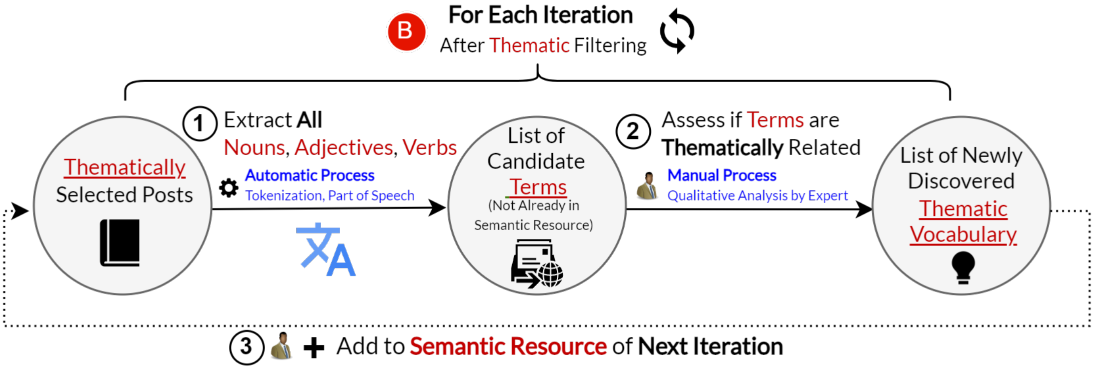

B
For Each Iteration
After Thematic Filtering
Thematically Selected Posts
1 Extract All Nouns, Adjectives, Verbs
Automatic Process
Tokenization, Part of Speech
List of Candidate Terms
(Not Already in Semantic Resource)
2 Assess if Terms are Thematically Related
Manual Process
Qualitative Analysis by Expert
List of Newly Discovered Thematic Vocabulary
3 + Add to Semantic Resource of Next Iteration

Figure 2.6: The Thematic Vocabulary Discovery Process

2.3.4 Dataset Preview and Iteration

Each iteration of the process yields a dataset of varying precision and completeness. At this point, the resulting dataset is previewed using a dedicated application (a web-based post viewer, that we will not present here) and evaluated by domain specialists. Their role is to assess a set of randomly selected posts (the more, the better) to determine whether there is too much noise (irrelevant posts), or too much silence (insufficient posts), and to identify any recurring types of posts that should be excluded or are missing.

Criteria for adding, removing, extending, or narrowing are then determined in collaboration with computer scientists, and a new iteration of the process begins using these refined criteria. This cycle continues until the final dataset is deemed satisfactory. This iterative and incremental feature is a cornerstone of our method, relying on an indefinite number of iterations, each improving upon the last. It can be likened to a *trial and error* process.

We will next demonstrate our methodology to a use case within our project: Tourism in the *French Basque Coast*.

2.4 Experimentation: Tourism in the *French Basque Coast*

For this experiment, we selected the social media platform X/Twitter. Indeed, it offered an API dedicated to researchers[2], allowing for the retrieval of up to 10 million tweets per month. This feature significantly facilitated the collection process by removing the requirement for complex web scraping setups. Unfortunately, this API was discontinued in April 2023 (Graham, 2023). Additionally, the volume of data available on this platform is both massive and diverse, with an estimated over 500 million tweets sent daily (Paolanti et al., 2021). We begin by defining the dataset (Subsection 2.4.1) and then proceed to apply our methodology (Subsection 2.4.2).

2.4.1 Tourism Dataset Definition

Let's start by defining the dataset. To define the *thematic dimension*, we rely on the *Thesaurus on Tourism and Leisure Activities* of the *World Tourism Organization* (WTO) (World Tourism Organization, 2002) [3], an extensive multilingual terminology (in French, English, and Spanish) standardizing and normalizing terms related to tourism and leisure activities. This resource covers roughly 1,300 concepts with 6,000 multilingual terms divided into 20 broad categories ranging from cultural heritage to tourism activities and the sociology of tourism.

For this experiment, we reduce the spatial extent of the data. We wish to gather content from only one specific sub-area which will serve as our *spatial dimension*: the *French Basque Coast*, broadly considered to be among the most touristic places in *France*.

Additionally, to showcase the full range of our dimensions, we will use the *temporal one* as well and focus solely on the *summer of 2019* season, summer being the season when visitors are usually the most active and 2019 the year before the Covid-19 pandemic, which could distort our analysis process. Given that our main dataset definition is quite narrow and highly focused, we do not set up a calibration dataset, as those are intended for very large and open datasets which may return a very large amount of data and significantly slow down the process.

[2] `https://developer.twitter.com/en/products/twitter-api/academic-research` (*discontinued in April 2023*)
[3] `https://www.e-unwto.org/doi/book/10.18111/9789284404551`

2.4.2 Methodology Implementation with X/Twitter

Table 2.3 illustrates the implementation of our methodology and the various iterations it underwent.

The process was implemented in *Python*, using the *Tweepy*[4] library for interacting with the X/Twitter API and *ElasticSearch*[5] for storing extracted tweets. We conducted three refinement iterations, focusing solely on tweets in French, English, and Spanish (*pre-filtering step*) posted during Summer 2019 (*temporal filtering*). To simplify the implementation process, we did not process media (e.g., pictures, videos, audios) associated with tweets.

Iteration 1

For geotagged tweets, we filtered them based on the bounding box of the study area, yielding approximately 7,000 tweets (Table 2.3, *Iteration 1*). These tweets were then thematically filtered using the complete WTO Thesaurus vocabulary in conjunction with IAM Entity Linker (Cossin et al., 2018), a dictionary-based approach for semantic annotation, resulting in 3,447 tweets. For non-geotagged tweets, we used multilingual toponyms (625 place names and points of interest) from *OpenStreetMap* (OSM) within the *French Basque Coast* area.

In the first iteration, we indiscriminately used all toponyms for filtering, which led to the retrieval of an excessive number of tweets (over 2.7 million). This prompted us to halt and refine the process in the subsequent iteration.

Iteration 2

Feedback from the first iteration guided us to blacklist professional or institutional accounts deemed irrelevant to our analysis in a second iteration (Table 2.3, *Iteration 2*). We also refined the list of toponyms, removing 46 overly common place names (e.g., *Golf Practice*, *Roman Bridge*), and narrowed down the tourism thesaurus to exclude certain branches considered irrelevant by our project domain experts (e.g., *Sociology of Tourism*, which included concepts like *Abortion*). This resulted in approximately 60,000 tweets by the end of the second iteration, with a significant number related to the G7, an important international event held in 2019 in the municipality of *Biarritz* within the region.

Iteration 3

A third iteration (Table 2.3, *Iteration 3*) was then conducted, introducing additional pre-filtering filters to blacklist G7-related keywords and hashtags (e.g., G7, G-7, #G7Biarritz), and a final pass on the tourism thesaurus ensured the inclusion of only concepts relevant to our project.

The final dataset, highlighted in green in Table 2.3, comprises 2,098 geotagged tweets and 25,281 tweets without geotagging (for a total of 27,379 tweets) from $\approx$ 15,000 users. 235 unique concepts were found in geotagged tweets and 458 in non-geotagged ones. We must now evaluate whether this dataset is relevant or not.

[4] https://www.tweepy.org

[5] https://www.elastic.co/elasticsearch

		Iteration 1		Iteration 2		Iteration 3	
		Geotagged Tweets	**Other Tweets**	**Geotagged Tweets**	**Other Tweets**	**Geotagged Tweets**	**Other Tweets**
Pre-Filtering	**Criteria**	**Languages**: *French*, *English*, *Spanish* **Blacklist**: *Retweets*, *Quotes*		+ **Blacklist**: *Professional Accounts*		+ **Blacklist**: *G7-Related* *Keywords* and *Hashtags*	
Temporal Filtering	**Criteria**	Summer 2019					
	Tweets	>1 billion tweets					
Spatial Filtering	**Criteria**	Basque Bounding Box	625 Basque OSM Places	Basque Bounding Box	579 Basque OSM Places	Basque Bounding Box	550 Basque OSM Places
	Tweets	7,003 tweets	>2,700,000 tweets	6,689 tweets	148,860 tweets	6,127 tweets	59,878 tweets
Thematic Filtering	**Criteria**	Full WTO Thesaurus	**Cancelled** (*too many tweets*)	Refined WTO Thesaurus		More Refined WTO Thesaurus	
	Tweets	3,447 tweets		2,390 tweets	59,968 tweets	**2,098 tweets**	**25,281 tweets**
Quantitative Statistics	**Hashtags (#)**	3,750 hashtags		3,620 hashtags	44,411 hashtags	3,341 hashtags	24,263 hashtags
	Users (@)	1,112 users		865 users	30,126 users	796 users	14,114 users
	Unique Places	32 locations		31 locations	194 locations	31 locations	184 locations
	Unique Concepts	462 concepts		245 concepts	540 concepts	235 concepts	458 concepts
	Top Concepts	*TourOperators*, *Beaches*, *Summer*, *Holiday*, *Days*		*Beacher*, *Summer*, *Holiday*, *Days*, *Ocean*	*Days*, *Town*, *Resident*, *Summer*, *Fetes*	*Beacher*, *Summer*, *Holiday*, *Days*, *Ocean*	*Fetes*, *Days*, *Beaches*, *Summer*, *Bullfight*

Table 2.3: Application of the Data Collection Methodology using our Tourism Dataset Requirements

2.5 Evaluation

We will now evaluate the collected dataset to assess the effectiveness of our methodology. We start by defining the quantitative and qualitative evaluation metrics we use (see Subsection 2.5.1) and then proceed to apply them to our resulting dataset (see Subsection 2.5.2 and Subsection 2.5.3).

2.5.1 Evaluation Metrics

We initially relied on quantitative statistics to evaluate the collected data (refer to Table 2.3). This involved evaluating the total number of tweets, the amount of unique users, locations, and the range of hashtags retrieved. Hashtags are metadata tags associated with a theme and are widely used on X/Twitter and many other social media platforms. Their inclusion in our analysis was therefore deemed valuable. Furthermore, we examined the number of concepts from the *World Tourism Organization Thesaurus on Tourism and Leisure Activities* (World Tourism Organization, 2002) that were employed to thematically filter tweets and identified those that were most prevalent. These quantitative metrics demonstrated the feasibility of our methodology with the social media platform X/Twitter, ensuring a substantial collection of tweets with minimal silence or excessive noise. The next step is to ascertain the accuracy and relevance of these collected tweets. To analyze and evaluate the quality of the dataset produced, we selected several evaluation metrics with varying degrees of granularity.

Firstly, we aim to evaluate the effectiveness of our thematic filtering process. Our reference points are the context annotations generated by X/Twitter (Elias, 2022). These annotations are contextual labels X/Twitter automatically assigns to tweets based on their content[6]. Although the specific methodology X/Twitter uses for these annotations is not disclosed, among them is a label for *Travel* (from our review, this is the only label related to tourism). Upon review, we found these annotations to be usually accurate, yet many relevant tweets lacked these labels (e.g., X/Twitter annotates accurately but does not annotate sufficiently), making it challenging to construct a large dataset relying solely on them. Other existing works have used these annotations as baselines (Gerosa and Ceinar, 2022; Oliveira et al., 2023), we therefore consider them reliable. Thus, it appeared pertinent to calculate the proportion of tweets identified as related to travel by X/Twitter that our system selects. The objective is to capture most of what X/Twitter tags as tourism-related and also to identify additional relevant tweets, thereby compiling a more substantial dataset for use.

Secondly, it is essential to assess whether the tweets not tagged as "*Travel*" that we incorporate are pertinent or merely noise. To this end, we conduct a qualitative analysis of the resulting dataset, focusing on a detailed examination of the content of collected tweets. We will assess the accuracy of a random subset of tweets at each iteration of our method. Several tourism experts will manually review their content to determine whether it pertains to tourism in the *French Basque Coast*. This qualitative metric is designed to highlight the impact of the different iterations on enhancing accuracy and to evaluate the overall quality of the dataset. Let's start with the quantitative analysis of X's context annotations.

[6] https://developer.twitter.com/en/docs/twitter-api/annotations/overview

2.5.2 Quantitative Analysis: Comparison with X's Context Annotations

Figure 2.7 shows the different sets of tweets of the last iteration (we use the union of both geotagged and non-geotagged ones).

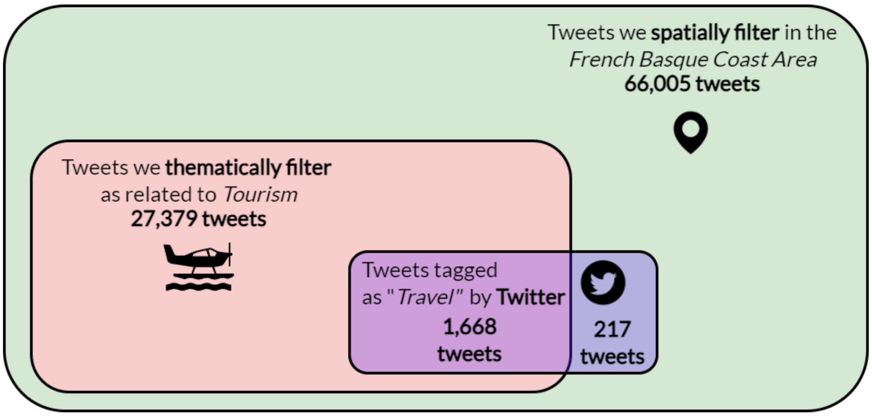

Tweets we **spatially filter** in the
French Basque Coast Area
66,005 tweets
Tweets we **thematically filter**
as related to *Tourism*
27,379 tweets
Tweets tagged
as "*Travel*" by **Twitter**
1,668 tweets
217 tweets

Figure 2.7: Number of Collected Tweets After Spatial and Thematic Filtering Steps During the Third Iteration Compared with Tweets Annotated as *Travel* by X/Twitter

Among these sets of tweets, we focus on the proportion annotated with the *Travel* context annotations by X/Twitter. We observe that within the tweets from the study area (66,005 tweets), the majority of those tagged as *Travel* by X/Twitter are selected by our system (1,668 selected, 217 excluded, see Figure 2.7), resulting in a recall of 0.88 on *Travel* annotated tweets (Recall $= \frac{\text{True Positives}}{\text{True Positives+False Negatives}} = \frac{1668}{1668+217} \approx 0.88$), which is quite high.

Additionally, we identify a significant number of selected tweets ($27,379 - 1,668 = 25,711$ tweets) not tagged by X/Twitter but selected by our process. This suggests that our methodology detects many more potentially relevant tweets. This outcome is seen as a positive aspect, indicating that X's context annotations may be missing from a considerable number of tourism-related tweets. The use of the thesaurus of tourism likely contributes to this result. Regarding the 217 tweets tagged as *Travel* by X/Twitter but excluded by our system, they were qualitatively reviewed by domain experts to assess if they were actually relevant or not. Approximately 76% were relevant but missed by our system due to containing highly misspelled touristic vocabulary or vocabulary not included in the thesaurus of tourism (165 tweets out of 217). Therefore, if we consider only tweets that should have been selected, our system exhibits an actual recall of 0.91 ($\frac{1668}{1668+165} \approx 0.91$), which is considered very high.

To summarize, out of the 27,379 tweets identified as related to tourism by our methodology, only 6,607 had X/Twitter context annotations (e.g., any context annotations, not only the *Travel* one), including 1,668 with the *Travel* annotation. This indicates that only about 6% of the tweets we collected were recognized as relating to tourism by X/Twitter ($\frac{1668}{27379} \times 100 \approx 6\%$). The next step is to determine whether the additional non-annotated tweets collected by our methodology are merely noise or if they represent additional tourism-related content that X/Twitter has not annotated.

2.5.3 Qualitative Analysis: Dataset Accuracy

For this purpose, a qualitative analysis is set up. 100, 50, and 20 random tweets are extracted from the datasets at different stages of the methodology and manually evaluated by experts to determine whether they have been correctly or wrongfully selected. This type of accuracy-based evaluation is widely used in the literature (Hossin and Sulaiman, 2015). Table 2.4 shows the result of this process at different stages of the method.

		Iteration 1		Iteration 2		Iteration 3	
		Geotagged	**Others**	**Geotagged**	**Others**	**Geotagged**	**Others**
Accuracy	(@ 20)	0.75	🟥	0.60	0.30	**0.83**	**0.72**
	(@ 50)	0.64		0.60	0.30	0.77	0.74
	(@ 100)	0.52		0.59	0.35	**0.74** (κ 0.74)	**0.65** (κ 0.48)

Table 2.4: Qualitative Analysis of the Results at Different Steps of the Methodology

Table 2.4 displays the accuracy measure for 20, 50, and 100 tweets, randomly selected at the end of the filtering flows (in our case, post-thematic filtering) from both sets of tweets, and subsequently manually evaluated by two experts. The question asked was "*Do you consider this tweet as related to Tourism in the French Basque Coast?*".

For now, let's focus on the third and last iteration. The accuracy measure @ 100 is complemented by the corresponding Cohen's Kappa Coefficient (κ) (Cohen, 1960), which gauges the level of concordance between the two experts to counterbalance the evaluations' subjectivity. The results reveal a mean accuracy ranging from 0.83 (@ 20) to 0.74 (@ 100) for geotagged tweets and 0.72 (@ 20) to 0.65 (@ 100) for other tweets. That means, by extrapolation, potentially 65% to 83% of the tweets not selected by X/Twitter's annotation system could indeed be pertinent to the domain of tourism, corroborating our hypothesis that X/Twitter is not annotating sufficient content.

Here is an example of a tweet detected by our system but overlooked by X/Twitter's annotation mechanism (translated from French):

> *"What a pleasure to **discover** this beautiful Saint Vincent d'Hendaye **church** while going to the market, So many lovely surprises! In **Basque Country** (**Hendaye**)."*

We also observe an accuracy @ 100 starting at 0.52 in the first iteration for geotagged tweets, which increases to 0.59 and then to 0.74 in iterations 2 and 3, clearly highlighting the effect of filter refinement and the feedback loop between each iteration. Overall, experts seem to agree on the outcome, as indicated by the relatively high Kappa (κ) scores in the last iteration (0.74 corresponds to a strong agreement and 0.48 to a moderate agreement). The accuracy for non-geotagged tweets is slightly lower but follows a similar increasing trend. The final dataset accuracy is acceptable, but it could have been further improved with more iterations, but we limited ourselves to three for this experiment.

To go further, we tracked the same sample of incorrect tweets from the first iteration throughout the methodology to observe what proportion would be removed in subsequent ones. We used the set of tweets from the first iteration's @ 100 accuracy measurement. The accuracy is average (0.52), meaning out of 100 tweets evaluated, 48 were incorrectly selected. The second iteration removed

27 of them, leaving 21 out of 48 remaining. Finally, the last iteration removed an additional 6, leaving 15 out of 48 remaining. This is consistent with the accuracy calculated on random samples previously (≈ 70% of tweets are correct). Below are examples of tweets incorrectly selected in iteration 1 that our method's feedback loop managed to eliminate in subsequent iterations:

1. *Okay, so experts, what did you think about yesterday's match in Bayonne? I almost couldn't watch it.* Selected in Iteration 1 due to *Match*, excluded in Iteration 2, 3 with the thesaurus vocabulary refinement (removal of *SportsCompetitions* concept branch).

2. *As part of the preparation for #G7Biarritz, I will have the honor of speaking during the Ocean, Our Future Forum on the relationship between oceans and public health.* Selected in Iteration 1, 2 due to *Ocean*, excluded in Iteration 3 due to blacklist of hashtag *#G7Biarritz*.

In summary, our methodology has demonstrated its effectiveness in collecting datasets that are relevant and comprehensive, with limited noise in the tourism domain. Its application to build a dataset in another domain of application (the domain of *local public policies*) and using another data source (municipality review platforms) will be showcased in Chapter 6. Let's now explore some perspectives to extend this methodology.

2.6 Summary and Perspectives

To conclude this chapter, here we proposed a novel, generic, and iterative methodology for building thematic datasets from social media. The objective is to move away from various ad-hoc collection processes and propose a robust, formalized methodology to build focused datasets from social media. This approach is based on several dimensions (spatial, temporal, and thematic) and operates in an iterative and incremental mode. The end-user is involved in assessing the produced datasets, and there is a mechanism of feedback loops to adjust the filtering and obtain datasets that do not have too much noise or silence. This contribution (Contribution 1) is positioned in the *Web and Social Media Search* research field. It addresses the challenge of constructing accurate and representative datasets from social media and has been experimented with by building a dataset about tourism in the *French Basque Coast*. The latter was evaluated using both quantitative (statistics, comparison with X/Twitter annotations) and qualitative (accuracy by domain experts) metrics.

However, it currently has limitations and could be improved in many ways. We have identified three main limitations and propose perspectives to alleviate them.

Firstly, the applicability of our methodology is primarily confined to social sites based on textual posts. This includes traditional social media platforms (e.g., *X/Twitter, Facebook*), review platforms (e.g., *TripAdvisor, Google Reviews*), or discussion sites (e.g., *forums*). Given that most social media are structured around textual content, our methodology should apply to a wide range of platforms. However, it may not be well-suited for other types of platforms, such as image or video-sharing platforms and blogs (blogs tend to have longer and deeper texts compared to social media). Additionally, the methodology we propose can be applied to the textual metadata associated with or extracted from multimedia content but not the media themselves. As discussed in Chapter 1, multimedia data are not the primary focus of our work, due to the distinct set of technologies required for their processing; therefore, we did not investigate this limitation.

Secondly, the implementation of the collection methodology we experimented with tourism professionals relies on several heterogeneous and non-integrated software modules, which can be a barrier for use with non-specialist users. It would be interesting to propose a generic software platform implementing this methodology for the most commonly used social media platforms (e.g., *X/Twitter, Facebook, Instagram*), dedicated to non-computer-scientist users. This would allow each project not to have to implement the methodology itself. This work was not carried out in the framework of this thesis because it is mainly engineering-related and not a research task.

Thirdly, we filter according to spatial, temporal, and thematic dimensions. We noticed that various types of entities, such as *people*, *organizations*, *currencies*, *amounts*, *temperatures*, etc., are also present in social media posts. Currently, we either do not leverage them (e.g., people, amounts, temperatures) or assimilate them as themes (e.g., organizations, currencies), and we would like to exploit them too. For example, one could imagine adding a weather dimension that would filter only tweets related to specific weather conditions. This would, for instance, be useful when the methodology is applied to the tourism domain to determine visitor activity during sunny or rainy conditions. Similarly, in the healthcare domain, leveraging named entities like *temperatures*, *amounts* (such as dosage information), and *organizations* (such as drug manufacturer) to construct a dataset could help build relevant social media dataset related to public health monitoring. To achieve this, we are thinking of extending the collection methodology further by extending our dimensions and by taking inspiration from the 5W1H dimensions. The 5W1H (Wang et al., 2010a) is a framework widely used in problem-solving and question-answering, based on the six interrogative words: *Who, What, When, Where, Why, How*. We will develop this perspective further in the general perspectives section at the end of the manuscript (refer to Section 7.2).

Let's now move to the next chapter, where we will study how we process the data collected in this step to extract fine knowledge from posts (*Transform* phase in Figure 1.5).

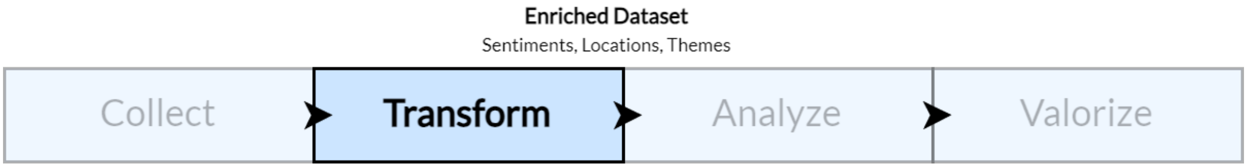
Chapter 3

Transform

Optimal Strategies for the Multilingual Analysis of Social Media Content in the Tourism Domain

> *"It is vain to do with more what can be done with fewer."*
>
> — William of Ockham, English Franciscan Friar and Philosopher

Processing vast volumes of social media data can be challenging (Maynard et al., 2012), especially when it comes to extracting structured knowledge from unstructured text, like posts on social media. This process is usually done through the use of Natural Language Processing (NLP) techniques. This chapter, associated with the *Transform* phase of the APs Framework (see Figure 1.5), attempts to address several NLP-related research challenges. It focuses on transforming unstructured text data into structured knowledge while dealing with the various complexities associated with social media posts, namely, their brevity and informal nature, the presence of grammatical errors, and special characters (hashtags, emojis, URLs, etc.). Additionally, a significant challenge is the multilingual nature of these texts.



We begin by introducing the main challenges in processing this kind of data (Section 3.1), then proceed to review existing NLP techniques, including both rule-based and deep learning-based ones, in three common knowledge extraction tasks: Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction, along with existing training resources (Section 3.2). Due to a lack of existing multilingual training resources in the domain of tourism, we propose and describe the process of creating a novel, manually annotated training dataset based on the data collected in Chapter 2 (Section 3.3, Contribution 2.1). This dataset serves as a basis for a comparative study (Section 3.4) between various techniques and language models to determine which ones are the best for these three knowledge extraction tasks in the domain of tourism. Additionally, for each deep learning technique, we seek to determine the tipping point at which annotating more data does not yield significantly better results (Section 3.5, Contribution 2.2) and discuss potential limitations (Section 3.6). Finally, we propose future extensions for our work (Section 3.7). The contribution introduced in this chapter has been published at the following national conference:

- M. Masson, R. Agerri, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose. (2023). Optimal Strategies for the Multidimensional Analysis of Multilingual Content from Social Media. *Proceedings of the 42nd Conference on Computer Science for Organizations and Information and Decision Systems (INFORSID 2024)* (Nancy, France).

3.1 Introduction: Extracting Knowledge from Unstructured Social Media Posts

Social media platforms have become essential channels for sharing opinions and experiences about tourism practices and itineraries. Consequently, tourism stakeholders (e.g., tourism offices, travel agencies, boards of touristic municipalities) often delegate the task of knowledge extraction to specialists, who rely on Natural Language Processing (NLP) techniques. NLP offers a powerful set of techniques for processing and analyzing text data and is often used to address three common knowledge extraction tasks (from now on, when we refer to "*the tasks*", it will be the following ones): Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction (Rosenthal et al., 2015; Liu et al., 2022; Fu et al., 2020).

3.1.1 Sentiment Analysis

Sentiment Analysis (Serrano-Guerrero et al., 2015) aims to determine the emotional tone behind a text and is used to gain an understanding of the attitudes, opinions, and emotions expressed within it. This task can involve classifying the polarity (e.g., *positive*, *negative*, *neutral*, etc.) of a given text at the document, sentence, or aspect level. In this work, we will focus on text-level classification, thereby making this task a text classification one (Gasparetto et al., 2022). Below is an example of Sentiment Analysis applied to a touristic post.

Positive { Yesterday, we went swimming at the Grande Plage beach in Biarritz, it was an **amazing** experience! The weather was **incredibly** sunny. **:)** #holidays #sun #btz

Note that another type of Sentiment Analysis exists: Aspect-based Sentiment Analysis (Nazir et al., 2020; Liu et al., 2020). It is a specialized form of Sentiment Analysis that focuses on identifying opinions or emotions and linking them with specific features of a topic within a text. It goes beyond general sentiment evaluation by pinpointing and evaluating the sentiments associated with each particular feature. Aspect-based Sentiment Analysis differs from regular Sentiment Analysis and will not be covered in this thesis.

3.1.2 Named Entity Recognition (NER) for Locations

Named Entity Recognition (NER) for Locations (Sun et al., 2018) seeks to identify and classify location named entities mentioned in texts into categories such as countries, municipalities, rivers, and mountains. Other types of Named Entity Recognition also exist to identify persons, organizations, dates, etc. (Goyal et al., 2018), but here we focus on locations. This is a sequence labeling task, also referred to as a token classification task. See the example below.

Yesterday, we went swimming at the $\overbrace{\text{Grande Plage}}^{\text{Location}}$ beach in $\overbrace{\text{Biarritz}}^{\text{Location}}$, it was an amazing experience! The weather was incredibly sunny. :) #holidays #sun # $\underbrace{\text{btz}}_{\text{Location}}$

3.1.3 Fine-grained Thematic Concept Extraction

Fine-grained Thematic Concept Extraction (Zhang, 2003) involves extracting fine-grained domain-specific concepts from text. These concepts are mapped to domain-specific semantic resources such as dictionaries, thesaurus, or ontologies. This is also a sequence labeling task. Below is an example of a tweet annotated using concepts from the *World Tourism Organization Thesaurus on Tourism and Leisure Activities* (World Tourism Organization, 2002).

Yesterday, we went $\overbrace{\text{swimming}}^{\text{Sports::WaterSports::Swimming}}$ at the Grande Plage $\overbrace{\text{beach}}^{\text{NaturalResources:Beaches}}$ in Biarritz, it was an amazing experience! The $\underbrace{\text{weather}}_{\text{NaturalResources::Weather}}$ was incredibly $\underbrace{\text{sunny}}_{\text{ClimaticFactors::Sun}}$. :) # $\underbrace{\text{holidays}}_{\text{VisitorFlows::Holidays}}$ # $\underbrace{\text{sun}}_{\text{ClimaticFactors::Sun}}$ #btz

Recently, NLP techniques based on deep learning and language models that can be used for these tasks have emerged (Birhane et al., 2023). These techniques offer several advantages over traditional rule-based approaches. Deep learning-based approaches can adapt to changing language patterns and structures (Min et al., 2023), ensuring a more dynamic and up-to-date analysis. Additionally, they can handle vast amounts of multilingual data efficiently. However, to achieve optimal results in domain-specific applications, language models must be fine-tuned. Fine-tuning is a process where a pre-trained model, initially trained on a large general dataset, is further trained on a smaller, domain-specific dataset to adapt its understanding and improve performance on tasks relevant to that domain (Iman et al., 2023).

Consequently, researchers often face two challenges: (1) **determining which NLP technique is most suitable for a given domain** (e.g., which technique, which language models, etc.), and (2) **determining how many domain-specific examples are necessary to achieve competitive NLP results**. As annotating datasets is both costly and time-consuming, researchers strive to keep the annotation work to a minimum while maintaining high-quality results.

We will now review existing NLP techniques for the three tasks as well as existing training resources for those based on deep learning processes.

3.2 Related Work: Natural Language Processing for Information Extraction

We have divided this related work section into five parts. We focus on the three common tasks evoked before, namely Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction. We start by reviewing traditional rule-based approaches (Subsection 3.2.1), then we move on to deep learning-based ones based on language models (Subsection 3.2.2) like fine-tuning (Subsection 3.2.3), few-shot prompting (Subsection 3.2.4), and cross-lingual transfer (Subsection 3.2.5). Lastly, we review existing training resources available for these three tasks (Subsection 3.2.6). In this section, we focus on the Natural Language Understanding (NLU) component of NLP (Khurana et al., 2023) (as opposed to Natural Language Generation), as it is what interests us in the context of our project.

3.2.1 Rule-based Techniques

Rule-based approaches are among the earliest techniques employed to handle text classification and sequence labeling tasks in NLP (Bajwa and Choudhary, 2006). This can be observed in the relative age of the references in this section. These approaches rely on a set of predefined rules or patterns designed to interpret and analyze text data. Various techniques are used (lexicon-based, pattern-based, grammar-based, and semantic-based), as highlighted in Table 3.1.

Lexicon-based Techniques

For Sentiment Analysis, rule-based systems typically employ a lexicon of sentiment-related words and phrases (Mohammad and Turney, 2013) (see Table 3.1, *Lexicon-based*) associated with their polarity and intensity scores to gauge the overall sentiment of a text. Notable examples include VADER (*Valence Aware Dictionary and sEntiment Reasoner*) (Hutto and Gilbert, 2014), which is specially designed for sentiments expressed in social media contexts; AFINN (Nielsen, 2017), an approach that weights words' polarities between +5 and -5; and SentiWordNet (Ohana and Tierney, 2009), an extension of WordNet (Miller, 1995) that associates terms with sentiment polarity.

In NER, particularly for location extraction, rule-based approaches rely on gazetteers or lists of place names to classify named entities within a text (Zhang, 2013), for example *OpenStreetMap*[1], *BD TOPO*[2] or *DBpedia Spotlight*[3] (Mendes et al., 2011).

For Fine-grained Thematic Concept Extraction, these approaches often use semantic resources such as dictionaries, serving as the lexicon, for example in the medical domain (Liu et al., 2012).

The main advantages of lexicon-based approaches include their simplicity in implementation and the interpretability of the results produced (Gehrmann et al., 2017). However, these approaches require to build a lexicon. While some terms may be universal (e.g., the meanings of emojis, certain place names), lexicon-based approaches require adaptation for each language. Additionally, these approaches often miss nuances in texts, such as negation, and do not consider misspelled terms, which is a frequent issue on social media (Clark and Araki, 2011).

[1] https://www.openstreetmap.org

[2] https://geoservices.ign.fr/bdtopo

[3] https://www.dbpedia-spotlight.org

Technique	**Main NLU Tasks**	**Strengths**	**Weaknesses**	**Examples**
Lexicon-based	*Sentiment Analysis, Named Entity Recognition, Concept Extraction*	Straightforward to implement, easily understandable	Require a lexicon, constrained by lexicon size, ignores context and grammatical structures	• Nielsen (2017) (*Sentiment Analysis*) • Hutto and Gilbert (2014) (*Sentiment Analysis*) • Ohana and Tierney (2009) (*Sentiment Analysis*) • Zhang (2013) (*Named Entity Recognition*) • Bodenreider (2004) (*Concept Extraction*)
Pattern-based	*Sentiment Analysis, Named Entity Recognition, Concept Extraction*	Mostly accurate for well-defined patterns	Misses variations not covered by the patterns, require correctly formatted sentences	• Diamantini et al. (2016) (*Sentiment Analysis*) • Ghag and Shah (2016) (*Sentiment Analysis*) • Ratinov and Roth (2009) (*Named Entity Recognition*) • Kitani et al. (1994) (*Concept Extraction*)
Syntax-based and Grammar-based	*Dependency Parsing, Part-of-Speech (PoS) Tagging*	Leverage linguistic structures for deeper analysis	Complex to maintain, especially in multilingual contexts	• Dozat and Manning (2016) (*Parsing*) • Socher et al. (2013a) (*Parsing*) • Manning et al. (2014) (*Parsing, Part of Speech tagging*) • Banko and Moore (2004) (*Part of Speech tagging*)
Semantic-based	*Semantic Analysis, Concept Extraction, Sentiment Analysis, Disambiguation*	Can understand nuanced meanings and relationships between terms	Requires comprehensive semantic knowledge bases, computationally intensive	• He et al. (2017) (*Semantic Analysis*) • Màrquez et al. (2008) (*Semantic Analysis*) • Kok and Domingos (2008) (*Semantic Analysis*) • Gaizauskas and Humphreys (1997) (*Semantic Analysis*) • Saif et al. (2012) (*Sentiment Analysis*) • Al-Harbi et al. (2017) (*Disambiguation*)

Table 3.1: Summary of Common Rule-based Approaches in Natural Language Processing

Pattern-based Techniques

To overcome the limitations of lexicon-based techniques, they are frequently combined with pattern-based techniques (see Table 3.1, *Pattern-based*). Such techniques employ pattern matching, often using keywords as entry points.

For instance, when it comes to Sentiment Analysis, these systems usually leverage sophisticated linguistic features, including negation handling (Diamantini et al., 2016; Ghag and Shah, 2016), to improve accuracy. Notably, the introduction of a negation word before a positive adjective can reverse the sentiment conveyed by the phrase.

Similarly, in the context of NER, pattern-based techniques can significantly refine precision (Ratinov and Roth, 2009) by recognizing common patterns associated with locations (Brando et al., 2016). Among the notable tools for NER that use pattern recognition are *mXS* (Nouvel, 2012), *PERDIDO* (Moncla and Gaio, 2023), *TEXTOMAP* (Brun et al., 2015), *CasEN* (Friburger and Maurel, 2004), and *Carte à Carte* (Dominguès and Eshkol-Taravella, 2015).

While pattern-based techniques can be an improvement over purely lexicon-dependent ones, they are not without their limitations. A significant limitation is the necessity to update and extend the pattern database regularly to reflect new linguistic trends and accommodate variations across languages. This task is both time-consuming and complex, as it requires a deep comprehension of language nuances and the specific domains where they are applied. Moreover, pattern-based systems may still encounter difficulties in understanding the most nuanced and context-dependent expressions, where sentiments or meanings are not overtly conveyed through identifiable patterns.

Syntax-based and Grammar-based Techniques

Syntax-based and grammar-based techniques (see Table 3.1, *Syntax-based and Grammar-based*) use the structural aspects of language to improve text analysis. They are used in tasks such as parsing (Dozat and Manning, 2016; Socher et al., 2013a) and part-of-speech (PoS) tagging (Manning et al., 2014; Banko and Moore, 2004).

These techniques analyze the grammatical structure of sentences to identify relationships between words. By applying rules that reflect the syntactic structure of a language, these systems can, for example, distinguish between noun and verb phrases, or identify subject-verb relationships.

The main challenge with these techniques is their complexity and the effort required to maintain and update the rules across languages in multilingual contexts. Developing these techniques requires extensive linguistic expertise, and scaling them is challenging due to the requirement for a detailed set of rules that takes into account the diversity and subtleties of human language. In addition, the effectiveness of these techniques is highly dependent on the quality of the input text, with short and poorly structured sentences, such as those found in social media, frequently leading to errors.

Semantic-based Techniques

Semantic-based techniques (see Table 3.1, *Semantic-based*) focus on understanding the meanings and relationships of words and phrases within the context of a larger text. Unlike previous techniques that primarily rely on word-level or phrase-level analysis, semantic techniques aim to understand the underlying concepts and themes, facilitating tasks such as Semantic Role Labeling (He et al., 2017; Màrquez et al., 2008; Kok and Domingos, 2008; Gaizauskas and Humphreys, 1997), Word Sense Disambiguation (Al-Harbi et al., 2017), Concept Extraction, and even Sentiment Analysis (Saif et al., 2012). These approaches often use semantic networks or ontologies, which map words to their meanings and relationships with other words, to interpret text at a deeper level.

The main advantage of semantic-based techniques are their ability to understand complex and nuanced meanings in texts, such as abstract concepts, metaphors, and specialized terminology. This is particularly useful in domains like medicine, where it can distinguish between different uses of terms (Ruch et al., 2001; Savova et al., 2008). However, semantic-based techniques are also associated with various challenges, including the requirement for extensive semantic knowledge bases, which are costly and time-consuming to build and maintain. Additionally, these techniques can require significant computational resources due to the complexity of parsing and understanding the relationships in semantic networks.

Recently there has been a noticeable shift from rule-based approaches to those based on deep learning (Kamath et al., 2019), as evidenced by the age of the references cited in this subsection. Several factors contribute to this transition, including the limited multilingual capabilities of rule-based methods, their maintenance complexity, and their tendency to overlook certain nuances in texts, especially in the context of social media texts that are short, informal, and with frequent grammar mistakes. We will now explore how deep learning has changed the field of NLP.

3.2.2 The Shift Towards Pre-Trained Language Models

To address the limitations of rule-based approaches, one of the most significant advancements in NLP has been the advent of pre-trained language models (LMs) based on the Transformer architecture (Vaswani et al., 2017). These models serve as foundational architectures trained on vast corpora, capturing a broad spectrum of linguistic structures, nuances, and knowledge (Manning et al., 2020; Min et al., 2023; Toporkov and Agerri, 2024). As a result, they offer a significant boost in performance and generalization for many NLP tasks, often in multilingual contexts. Here, we will focus on both types of language models based on the Transformer architecture:

- *Masked Language Models* (we refer to them as MLMs), use the *encoder* block of the Transformer (Vaswani et al., 2017). The learning objective of MLMs consists of learning to predict masked words from the surrounding context. Popular models include BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) or XLM-RoBERTa (Conneau et al., 2019).

- *Large Language Models* (we refer to them as LLMs) are text-to-text models based on both blocks (*encoder-decoder*) or only the *decoder* component of the Transformer. These models are generative and, while their most successful results have come in text generation tasks, they have also been started to use for discriminative tasks in few-shot settings (Chung et al., 2024; García-Ferrero et al., 2024; Sainz et al., 2024). Generative LLMs include the GPT (*Generative Pre-Trained Transformer*) series of models (Brown et al., 2020), Mistral (Jiang et al., 2023), LLaMA 2 (*Large Language Model Meta AI*) series (Touvron et al., 2023b) or Google's FlanT5 (Chung et al., 2024) to name but a few.

Table 3.2 presents a comparison of pre-trained language models, base versions of models were used in this comparison.

Name and Reference	Year	Parameters	Training Corpus	Licence	Language	Type
GPT 2 (Radford et al., 2019)	2019	1.5 billion	10 billion tokens	MIT	Multilingual	LLM
GPT 3 / GPT 3.5 (Brown et al., 2020)	2020	175 billion	300 billion tokens	Proprietary	Multilingual	LLM
GPT 4 (Achiam et al., 2023)	2023	Trillions	Unknown	Proprietary	Multilingual	LLM
BERT (Devlin et al., 2019)	2018	110 million	3.3 billion words	Apache 2.0	English	MLM
Multilingual BERT (Devlin et al., 2019)	2018	110 million	3.3 billion words	Apache 2.0	Multilingual	MLM
LaMDA (Thoppilan et al., 2022)	2022	137 billions	168 billion tokens	Proprietary	English	LLM
ERNIE 3.0 (Sun et al., 2021)	2021	260 billion	4 TB of text	Proprietary	Chinese	LLM
Jurassic-1 (Lieber et al., 2021)	2021	178 billion	300 billion tokens	Proprietary	Multilingual	LLM
XLM-RoBERTa (Conneau et al., 2019)	2019	125 million	2.5 TB of text	MIT	Multilingual	MLM
RoBERTa (Liu et al., 2019)	2019	125 million	160 GB of text	MIT	English	MLM
DistilBERT (Sanh, 2019)	2019	66 million	3.3 billion words	Apache 2.0	English	MLM
CamemBERT (Martin et al., 2020)	2019	110 million	138 GB of text	MIT	French	MLM
ALBERT (Lan et al., 2019)	2019	11 million	16 GB of text	Apache 2.0	English	MLM
FlauBERT (Le et al., 2020)	2019	137 million	Unknown	MIT	French	MLM
XLNet (Yang et al., 2019)	2019	110 million	33 billion words	Apache 2.0	English	MLM
FlanT5 (Raffel et al., 2020)	2019	220 million	800 GB of text	Apache 2.0	Multilingual	LLM
LLaMA 7B (Touvron et al., 2023a)	2022	7 billion	1.4 trillion tokens	Proprietary	English	LLM
LLaMA 2 7B (Touvron et al., 2023b)	2023	7 billion	2 trillion tokens	Proprietary	English	LLM
Mistral 7B (Jiang et al., 2023)	2023	7.3 billion	Unknown	Apache 2.0	Multilingual	LLM
StableLM 2 (Bellagente et al., 2024)	2024	1.6 billion	100 billion tokens	Apache 2.0	Multilingual	LLM
PaLM (Chowdhery et al., 2023)	2022	540 billion	768 billion tokens	Proprietary	Multilingual	LLM
PaLM 2 (Anil et al., 2023)	2022	340 billion	3.6 trillion tokens	Proprietary	Multilingual	LLM
Latxa (Etxaniz et al., 2024)	2024	7 billion	4.2 billion tokens	MIT	Basque	LLM

Table 3.2: Comparison of Existing Pre-Trained Language Models

We have highlighted in green the candidate models we will use in our experiments (refer to Section 3.4). For each type of model (e.g., both MLMs and LLMs), we have selected the current state-of-the-art models, widely recognized as the most efficient ones, publicly available at the time of this study.

The models in Table 3.2 are not directly trained for the three knowledge extraction tasks we are interested in. They are generative LLMs trained to predict the next token in a sequence or MLMs trained to predict masked words in texts. For them to apply to our three tasks, namely Sentiment Analysis, NER for Locations, and Fine-grained Thematic Concept Extraction, an additional step is required. This step requires a training corpus and can be carried out via different techniques. Let's present the first one: fine-tuning.

3.2.3 Fine-Tuning of Language Models

A popular approach in deep learning-based NLP is to fine-tune language models for domain-specific downstream tasks (see Table 3.3 and Figure 3.1). Fine-tuning is a process where a pre-trained model, initially trained on a large general dataset, is further trained on a smaller, domain-specific dataset to adapt its understanding and improve performance on tasks relevant to that domain. This results in altering the model weights to adapt it to the new task.

For instance, language models have been fine-tuned for text classification tasks, including spam detection in hotel reviews (Crawford and Khoshgoftaar, 2021) and Sentiment Analysis in touristic reviews (Enríquez et al., 2022; Vásquez et al., 2021) or reviews about sustainable transport (Serna et al., 2021), leading to improved accuracy.

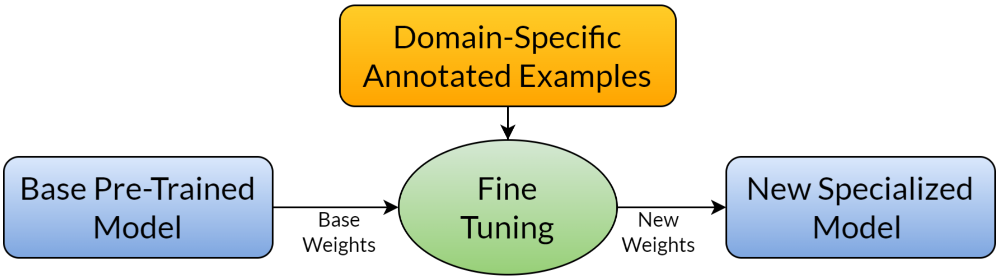


Figure 3.1: Overview of the Fine-Tuning Process

In NER, fine-tuning of language models has been employed to extract location information from tourism corpora (Bouabdallaoui et al., 2022; Cheng et al., 2020). Furthermore, language models have demonstrated promising results in Thematic Concept Extraction, such as identifying travel-related themes and topics from tourism texts (Chantrapornchai and Tunsakul, 2021).

Reference and Year	Objective	Label	Fine-Tuned Model(s)
Crawford and Khoshgoftaar (2021)	Spam Detection (Hotel Reviews)	Text	BERT
Enríquez et al. (2022)	Sentiment Analysis (Tourism Reviews)	Text	RoBERTa (RoBERTaESP)
Vásquez et al. (2021)	Sentiment Analysis (Tourism Reviews)	Text	BERT (BETO)
Serna et al. (2021)	Sentiment Analysis (Transport)	Text	XLM-RoBERTa
Landa and Agerri (2021)	Classification of Basque Users	Text	mBERT (BERTeus)
Bouabdallaoui et al. (2022)	Location Extraction (Touristic Corpus)	Token	BERT, RoBERTa, XLM-RoBERTa
Cheng et al. (2020)	Location Extraction (Touristic Corpus)	Token	BERT
Chantrapornchai and Tunsakul (2021)	Travel Themes Identification	Token	BERT
Chen et al. (2021)	Covid-19 Fake News Detection	Text	BERT, ALBERT, RoBERTa
Kumar et al. (2021)	Fake News Detection	Text	XLNet
Tripathy et al. (2022)	Cyberbullying Detection	Text	ALBERT

Table 3.3: Application of Language Model Fine-Tuning in Various Domains

One of the main limitations of fine-tuning is that the domain-specific dataset used in this process must be large enough and cover many different cases for the model to effectively learn the tasks (Devlin et al., 2019). Building and annotating this domain-specific dataset is often complex when there are no existing resources linked to the domain. It is therefore a costly and time-consuming process that researchers strive to avoid. As a result, novel training techniques that require only a few examples have appeared recently: few-shot prompting techniques.

3.2.4 Addressing the Lack of Domain-Specific Annotated Data with Few-Shot Techniques

Zero-shot and few-shot learning techniques (see Figure 3.2) have emerged as effective approaches to mitigate the requirement of manually annotated training data. Thus, instead of fine-tuning the pre-trained model's weights to a downstream task, prompting the language models in zero and few-shot settings allows to obtain competitive results in classification tasks (Kadam and Vaidya, 2020).

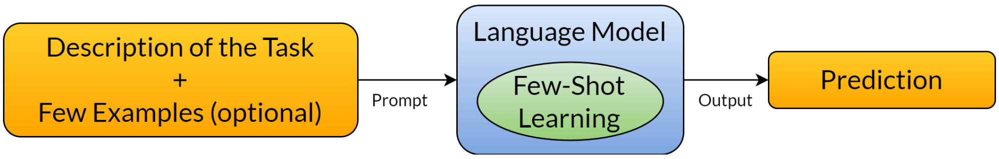


Figure 3.2: Overview of the Few-Shot Learning Process

Few-Shot with Generative Large Language Models (LLMs)

In the case of generative LLMs (such as GPT (Brown et al., 2020), Mistral (Jiang et al., 2023) or LLaMA 2 (Touvron et al., 2023b)), it is possible to apply them in zero-shot by simply describing the task to be carried out in natural language. Most commonly they are generation tasks, but they can also be prompted to perform text classification and sequence labeling tasks such as Sentiment Analysis and NER, respectively. Furthermore, sometimes adding a few examples may help as illustrated by the following 2-shot prompt for Sentiment Analysis.

```
You are an assistant that classifies sentiments of texts.
You must classify them as: positive, negative, or neutral.
Examples:
User: "We went to the beach yesterday, it was amazing!"
Assistant: positive
User: "So bad, it's raining today. Have to stay home ... :("
Assistant: negative
User: "Beautiful sun today"
Assistant: ...
```

Few-shot prompting is interesting, especially in scenarios in which domain-specific annotated data is rare and has been applied with promising results (Brown et al., 2020). However, results across domains are mixed. For example, studies have found that it can perform poorly in some domains, like the biomedical one (Moradi et al., 2021; Gutiérrez et al., 2022).

Few-Shot with Masked Language Models (MLMs)

Concerning text classification tasks, Pattern-Exploiting Training (PET) is a semi-supervised few-shot training approach that uses MLMs as the backbone. It combines the idea of providing the MLM with task descriptions in natural language and a cloze-style phrase generation approach to help the model understand the task (Schick and Schütze, 2021). For example, to classify movie reviews based on the predominant sentiment they express, the model would be prompted with the query: *The movie was* $\langle MASK \rangle$. The model would then try to predict the $\langle MASK \rangle$, choosing from options such as outstanding (*positive*) or terrible (*negative*).

More recently, SetFit (Tunstall et al., 2022) provides a prompt-free framework for few-shot fine-tuning of *Sentence Transformers*. It leverages *contrastive learning*, where only a small number of labeled examples are needed to fine-tune a pre-trained model. (Tunstall et al., 2022). SetFit attains high accuracy using minimal labeled data. For example, it requires just eight labeled examples per class on the customer reviews sentiment dataset to be competitive with fine-tuned RoBERTa-large (Liu et al., 2019) on the full training set of 3k examples (Tunstall et al., 2022).

Regarding sequence labeling tasks, several recent studies have also explored new approaches to replace complex templates used in few-shot prompting such as for NER (Wang et al., 2022; Ma et al., 2022). For example, EntLM (*Entity-oriented LM*) (Ma et al., 2022) aims to simplify the process of generating task-specific queries and reduce the reliance on manual template construction. EntLM currently represents the state-of-the-art for few-shot NER.

3.2.5 Cross-lingual Transfer Techniques

An alternative to few-shot prompting to address the lack of annotated data for NLP tasks on social media is the use of multilingual language models to perform data augmentation via machine translation or cross-lingual model-transfer.

In the first case, the idea is to translate into multiple languages the annotated training data from a source language and then use the translated versions to perform data augmentation during training. This technique has been tested for many sequence labeling (García-Ferrero et al., 2022, 2023; Yeginbergen and Agerri, 2024) and classification tasks (Artetxe et al., 2020, 2023) in various genres of text, including Sentiment Analysis in social media (Barriere and Balahur, 2020).

In the second case, the idea is to leverage the multilingual capabilities of some MLMs and LLMs to learn in a source language and predict in a different target language. Although interesting, cross-lingual transfer techniques are based on transferring *existing* annotations from (at least) a source to a given target language(s). However, our starting point is the absence, at the time of this study, of any publicly accessible, domain-specific annotated data in any language for touristic locations, fine-grained touristic concepts, and sentiment for the tourism domain. Let's now examine existing annotated resources for model training in other domains.

3.2.6 Existing Annotated Resources

While publicly available annotated data for the tourism domain is non-existent (paid datasets exist, but we will not cover them), there are several existing annotated corpora from other domains that could be used for experimentation, as highlighted in Table 3.4. Here, we compare a selection of existing annotated datasets along different criteria: (1) the source of data collection, (2) the types of annotations available, (3) the languages covered by the dataset, and (4) the method used for generating annotations (manual by humans, semi-automatic, or automatic).

For example, the *ESTER* corpus (Galliano et al., 2006) is a comprehensive collection of French radio transcripts, and *AnCora* (Taulé et al., 2008) is a multilevel annotated corpus (mostly from newspapers) for Catalan and Spanish. *FEW-NERD* (Ding et al., 2021) is an English corpus sourced from *Wikipedia*. These resources are annotated for NER (such as persons, locations, and organizations). Recently, multilingual NER corpora like *MultiCoNER* (Malmasi et al., 2022) have appeared, it is a multilingual (11 languages) dataset sourced from *Wikipedia*.

In terms of social media-specific resources, the *Broad Twitter Corpus* (BTC) (Derczynski et al., 2016) includes coarse-grained NER annotations, while *Sentiment140* (Go et al., 2009), *STS-Gold* (Saif et al., 2013), and many other datasets developed as part of shared evaluation tasks at *SemEval* (Nakov et al., 2013; Rosenthal et al., 2014, 2015; Nakov et al., 2016; Rosenthal et al., 2017), are used for Sentiment Analysis.

Dataset	Source	Annotations	Language	Type
CoNLL02 (T. K. Sang, 2002)	Newspaper	Named Entities	🇪🇸 🇳🇱	Manual
CoNLL03 (T. K. Sang and Meulder, 2003)	Newspaper	Named Entities	🇬🇧 🇩🇪	Manual
ESTER (Galliano et al., 2006)	Radio	Named Entities	🇫🇷	Manual
AnCora (Taulé et al., 2008)	Newspapers	Named Entities	🇪🇸	Semi-Auto
BTC (Derczynski et al., 2016)	X/Twitter	Named Entities	🇬🇧	Manual
CLUENER2020 (Xu et al., 2020)	Newspaper	Named Entities	🇨🇳	Semi-Auto
FEW-NERD (Ding et al., 2021)	Wikipedia	Named Entities	🇬🇧	Human
MultiCoNER (Malmasi et al., 2022)	Wikipedia	Named Entities	12 languages	Auto
Sent140 (Go et al., 2009)	X/Twitter	Sentiment	🇬🇧	Auto
IMDB (Maas et al., 2011)	Movie Reviews	Sentiment	🇬🇧	Auto
SST (Socher et al., 2013b)	Movie Reviews	Sentiment	🇬🇧	Human
STS-Gold (Saif et al., 2013)	X/Twitter	Sentiment	🇬🇧	Manual
GSC (Serna et al., 2021)	Transport Reviews	Sentiment	🇬🇧	Auto
SemEval (Nakov et al., 2013)	X/Twitter	Sentiment	🇬🇧	Manual
MultiWOZ (Budzianowski et al., 2018)	Humans	Dialogue States	🇬🇧	Manual
SNLI (Bowman et al., 2015)	Humans	Inference Pairs	🇬🇧	Manual
Heldugazte (Landa and Agerri, 2021)	X/Twitter	Formal/Informal	🇬🇧	Auto
SQuAD (Rajpurkar et al., 2018)	Wikipedia	Question Answering	🇬🇧	Manual
QuAC (Choi et al., 2018)	Wikipedia	Question Answering	🇬🇧	Manual
NQ (Kwiatkowski et al., 2019)	Google Search	Question Answering	🇬🇧	Semi-Auto

Table 3.4: Comparison of Existing Annotated Corpus

Other annotated datasets include the *Stanford Sentiment Treebank (SST)* (Socher et al., 2013b) and *IMDB* (Maas et al., 2011) based on automatically annotated movie reviews (for Sentiment Analysis). Additionally, corpora such as the *MultiWOZ* dialogue dataset (Budzianowski et al., 2018), the *Stanford NLI* dataset (Bowman et al., 2015) for text inference, and the *Heldugazte* corpus (Landa and Agerri, 2021), which assists in categorizing tweets as formal or informal, are noteworthy. Lastly, some corpora like the *Stanford Question Answering Dataset (SQuAD)* (Rajpurkar et al., 2018), *Question Answering in Context (QuAC)* (Choi et al., 2018), and *Natural Questions (NQ)* (Kwiatkowski et al., 2019) are used for question answering.

These datasets are extensive but broad and often focus on English only, therefore lacking the necessary contextual information relevant to the tourism domain. Most importantly, we could not find any public dataset annotated for Fine-grained Thematic Concept Extraction in the tourism domain. Taking this into account, we decided to build our own custom annotated dataset.

3.3 Dataset Building and Annotation

In this section, we describe the creation of a novel multilingual dataset consisting of tourism-related tweets annotated for three common NLP tasks for touristic applications: (1) Sentiment Analysis, (2) Named Entity Recognition for Locations, and (3) Fine-grained Thematic Concept Extraction (based on the *Thesaurus on Tourism and Leisure Activities of the World Tourism Organization* (World Tourism Organization, 2002)).

3.3.1 Data Collection

The raw dataset we use is the one collected from X/Twitter using our collection methodology, as described in Chapter 2. As a reminder, the dataset was defined within the following scope:

- *Spatial*: The *French Basque Coast* area, defined by spatial coordinates for geotagged tweets or a list of toponyms.
- *Temporal*: The summer of 2019, from June 21st to September 21st, based on the timestamps of the tweets.
- *Thematic*: The tourism domain, as defined by the *World Tourism Organization Thesaurus on Tourism and Leisure Activities*.

The final dataset consisted of 27,379 tweets, out of which we decided to select 2,961 tweets corresponding to 624 users for annotation and further use in our experiments. Several reasons motivate this choice:

- This subset of tweets (2,961 tweets) was manually verified to ensure that they were **both** *related to tourism* and *authored by visitors*, as opposed to tweets from tourism professionals or news outlets discussing tourism, among others. Additionally, in the future, we are looking to use these user annotations to train an automatic user classifier for the tourism domain.
- We have limited resources, budget, and time for annotation. Many massive annotated datasets rely on large teams or external crowdsourcing (such as *Amazon Mechanical Turk* (Jiang et al., 2022; Turcan and Mckeown, 2019; Lawson et al., 2010)) to annotate the data. We do not have these resources at our disposal.
- Focus on low-data techniques. We are looking to focus our experimentation on low-data techniques that allow us to obtain competitive results without requiring too many training examples.

The dataset is multilingual, comprising an unbalanced mix of tweets in English, French, and Spanish. This diversity reflects the actual use of social media in the *French Basque Coast* area.

	All Tweets	**French Tweets**	**English Tweets**	**Spanish Tweets**
Train	1,662 (503 users)	1,297 (391 users)	283 (129 users)	82 (32 users)
Dev	**619** (300 users)	450 (213 users)	99 (66 users)	70 (31 users)
Test	**680** (431 users)	401 (273 users)	102 (100 users)	177 (93 users)

Table 3.5: Breakdown of the Collected Dataset by Language – Tweets (Users)

Table 3.5 presents the language distribution within the dataset and the splits created for experimental purposes (60% for training, 20% for development, and 20% for testing). These splits were designed to maintain a balance in the number of users and languages represented in each set.

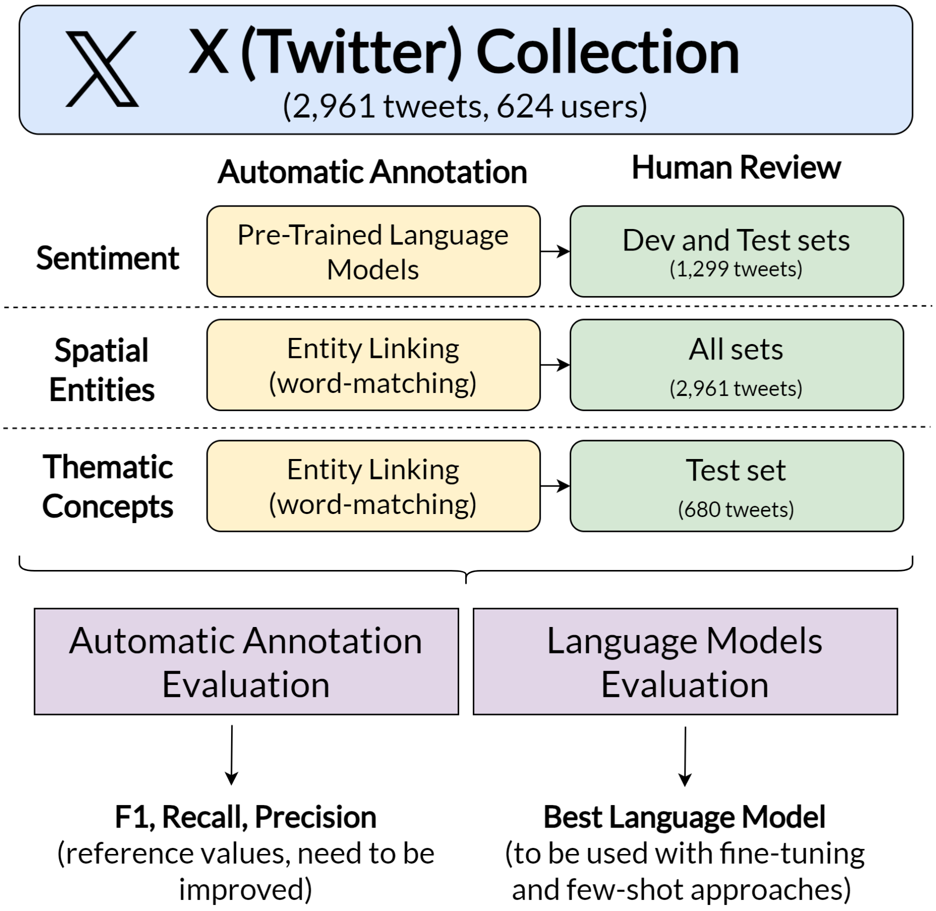

X (Twitter) Collection
(2,961 tweets, 624 users)
Automatic Annotation
Human Review
Sentiment
Pre-Trained Language Models
Dev and Test sets
(1,299 tweets)
Spatial Entities
Entity Linking (word-matching)
All sets
(2,961 tweets)
Thematic Concepts
Entity Linking (word-matching)
Test set
(680 tweets)
Automatic Annotation Evaluation
Language Models Evaluation
F1, Recall, Precision
(reference values, need to be improved)
Best Language Model
(to be used with fine-tuning and few-shot approaches)

Figure 3.3: Dataset Building and Annotation Process

The annotation process of those 2,961 tweets was carried out in a semi-automatic manner, following the procedure depicted in Figure 3.3. Let's get into details.

3.3.2 Sentiment Annotations

Firstly, to assist human annotators, the 1,299 tweets in the development and test splits underwent a process of automatic annotation using the five language models listed in Table 3.6.

Subsequently, they were manually reviewed. Each tweet was assigned to two annotators to evaluate the agreement (Cohen's kappa coefficient (Cohen, 1960)) and ensure the quality of the annotations. We achieved $\kappa = 0.79$ for French tweets, $\kappa = 0.75$ for Spanish tweets, and $\kappa = 0.67$ for English tweets, which corresponds to a strong agreement. Any disagreements were resolved through collaborative discussion.

The next step was to evaluate the performance of the five language models used to automatically label the tweets with respect to human annotations. Table 3.6 shows that XLM-T Sentiment, fine-tuned with multilingual Sentiment Analysis data from various domains different to tourism (Barbieri et al., 2022), outperformed, on average for the three languages, any other approach. Following this, we annotated the training split using XLM-T Sentiment.

Sentiment Models:	Barbieri et al. (2020)	Pérez et al. (2021)	Seethal (2023)	Hartmann et al. (2023)	Barbieri et al. (2022)
	French				
All Tweets	0.56	0.45	0.43	0.47	**0.82**
Positive Tweets	0.34	0.14	0.11	0.95	0.82
Negative Tweets	0.06	0.11	0.00	0.28	1.00
Neutral Tweets	0.97	0.97	1.00	0.00	0.74
	Spanish				
All Tweets	0.71	0.64	0.61	0.34	**0.83**
Positive Tweets	0.31	0.09	0.03	1.00	0.84
Negative Tweets	0.00	0.00	0.00	0.29	0.43
Neutral Tweets	1.00	1.00	0.98	0.00	0.87
	English				
All Tweets	**0.81**	**0.81**	0.71	0.66	0.80
Positive Tweets	0.75	0.75	0.59	1.00	0.72
Negative Tweets	0.75	0.50	0.75	0.50	0.75
Neutral Tweets	0.94	0.97	0.94	0.00	0.97

Table 3.6: Accuracy of Available Sentiment Language Models on Manually Annotated Test Data.

3.3.3 Locations and Fine-grained Thematic Concepts Annotations

Although Sentiment Analysis is a text classification task in which each tweet is assigned a polarity label, we are also interested in identifying locations and fine-grained thematic concepts relevant to the tourism domain. These two tasks are addressed as sequence labeling problems. Before experimenting with deep learning approaches, we implemented a basic rule-based *word-matching* (refer to Subsection 3.2.1, *Lexicon-based Strategies*) approach as a baseline for locations and fine-grained thematic concepts.

Locations were matched using 625 local toponyms extracted from *OpenStreetMap*[4] (*municipalities, POIs, landmarks, etc.*) while fine-grained thematic concepts were matched using their label and synonyms in the *Thesaurus on Tourism of Leisure Activities of the World Tourism Organization* (which contains 1,494 touristic concepts) (World Tourism Organization, 2002). Tweets preprocessing (lowercase, removal of URLs, hashtag splitting, decomposing hashtags to find concepts or toponyms in them) was performed to facilitate *word-matching*. We applied this algorithm to annotate the *train*, *dev*, and *test* splits. Automatic annotations were then manually corrected by human annotators, for locations in all *train*, *dev*, and *test* sets. For fine-grained thematic concepts, the *word-matching* algorithm detected 315 unique concept classes for the full dataset (out of the 1,494 concepts included in the WTO thesaurus), making it a highly fine-grained sequence labeling task. Due to this fact, we

[4] https://www.openstreetmap.org

only revised the fine-grained thematic concepts for the test set, as annotating 315 concept classes is a complex task requiring a large human effort.

Finally, inter-annotator agreement was calculated on a subset of 100 random tweets. For location entities: $\kappa = 0.91$ for exact matches, when all tokens forming an entity are precisely the same (e.g., New (B-LOC), - (I-LOC), York (I-LOC)). Here, B-LOC stands for **B**eginning-**Loc**ation and I-LOC stands for **I**ntermediate-**Loc**ation. On the other hand, $\kappa = 0.93$ for partial matches, when an entity is mostly recognized but has missing or extra tokens (e.g., New (B-LOC), - (O), York (O)). Both values indicate a near-perfect consensus.

The F1-score results of evaluating the *word-matching* baseline on the test set are reported in Table 3.7. These results will serve as a baseline to compare with the different supervised approaches in the experimental section.

Named Entity Recognition (NER) for Locations			
	Recall	**Precision**	**F1-score**
Location Exact Match	0.692	0.722	0,707
Location Partial Match	0.780	0.814	0,796
Fine-grained Thematic Concept Extraction			
	Recall	**Precision**	**F1-score**
Concept Exact Match	0.746	0.952	0,836
Concept Partial Match	0.747	0.953	0,837

Table 3.7: Performance of the *Word-Matching* Algorithm on both Sequence Labeling Tasks

From the results in Table 3.7, it can be observed that, for locations, results are not satisfying, particularly in terms of recall. However, the *word-matching* algorithm performs remarkably well on Fine-grained Thematic Concept Extraction, especially in terms of precision. Although this constitutes a rather strong baseline, the recall remains comparatively low, which means that many fine-grained thematic concepts remain undetected by the system. Furthermore, the *word-matching* algorithm is a rather rigid and static system, which we would ideally like to avoid for new domain-specific applications.

Thus, our main objective is now to establish whether deep learning supervised techniques based on multilingual language models can match or improve over the *word-matching* algorithm while keeping the amount of manual annotation to a minimum, especially for Fine-grained Thematic Concept Extraction. Thus, in addition to standard fine-tuning approaches, it is of particular interest investigating techniques based on few-shot prompting, where the aim is to generate competitive taggers using only a very small amount of labeled data.

The data used for experimentation is the one generated by manually revising the annotations for the three tasks, as described above. Table 3.8 provides a detailed description of the dataset including the total number of tweets (2,961) and the number of annotations per task. Note that for the sentiment column in Table 3.8, the colors and symbols used represent each type of polarity (- for negative annotations, + for positive and = for neutral). Appendix B shows an excerpt from the dataset, specifically a post annotated for the three tasks.

Set	Tweets	Locations	Concepts	Distribution of Sentiments
Train	1,662	4,030	3,841	787 (+) 191 (-) 684 (=)
Dev	619	1,419	1,337	271 (+) 82 (-) 266 (=)
Test	680	1,679	1,844	299 (+) 93 (-) 288 (=)
All	2,961	7,128	7,022	1,357 (+) 366 (-) 1,238 (=)

Table 3.8: Dataset Annotations (Following Human Review)

In summary, the goal is to establish how much labeled data do we actually require to obtain competitive performance by comparing the fine-tuning and few-shot approaches with respect to the *word-matching* algorithm, and which of these approaches is the most efficient in terms of performance and human effort.

3.4 Experimental Setup

Figure 3.4 provides an overview of our experimental setup with the models, sampling methods, and learning techniques used for each task. The experimentation is focused on the three tasks presented above: Sentiment Analysis (Subsection 3.4.1), NER for Locations and Fine-grained Thematic Concept Extraction (Subsection 3.4.2).

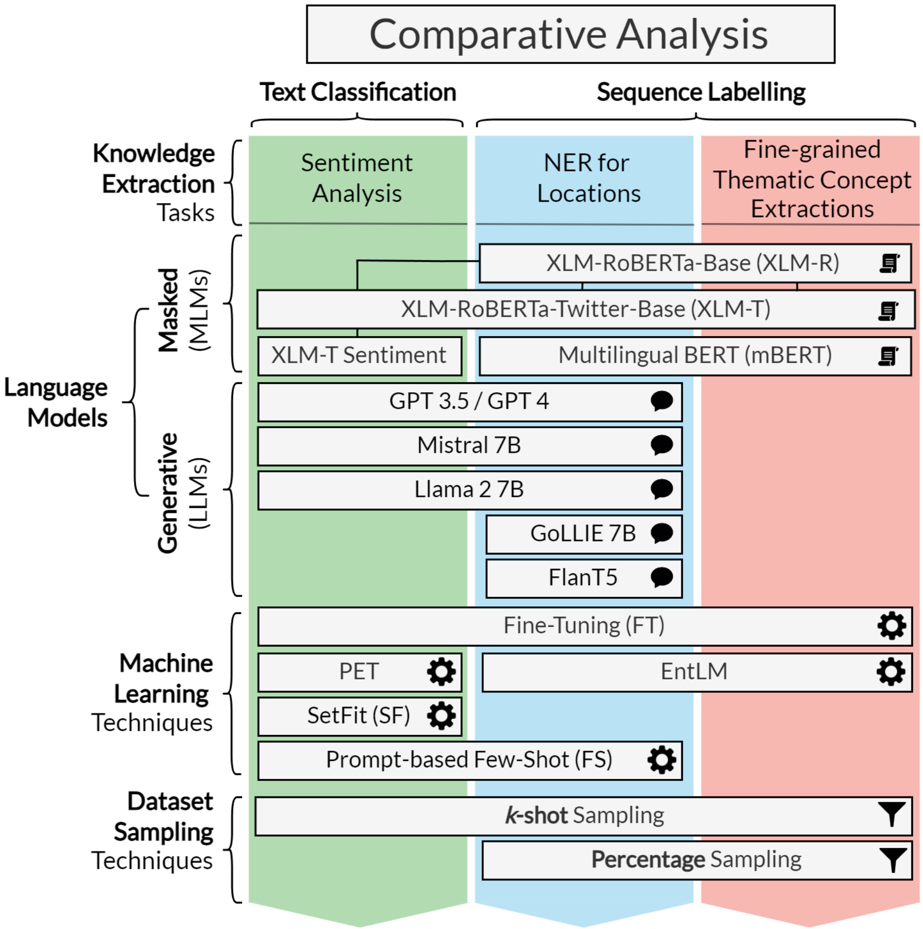


Figure 3.4: Experimental Setup of the Comparative Study

These experiments leverage the tourism dataset described in the previous section (*2,961 multilingual tweets including 1,662 for training*) and summarized in Table 3.8. The dataset is sampled using two different methods:

- *k-shot sampling*: In this technique, we selected a specific number of examples for each tweet or token label from the training set. We performed training or prompting using the following k-values: 5, 10, 20, 30, 40, 50, and 100 examples per label. For Sentiment Analysis, we used the PET k-shot sampling technique (Schick and Schütze, 2021) while for locations and fine-grained thematic concepts, we apply the EntLM k-shot technique with default parameters (Ma et al., 2022).
- *Percentage sampling*: For sequence labeling (locations and fine-grained thematic concepts) we also experiment with sampling on percentages of tweets rather than labels, as k-shot does. We successively used 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 90%, and 100% of the training set, while trying to maintain the original label classes distribution, including the O labels (the O label is assigned to tokens that are neither location nor thematic entities), which the k-shot sampling technique does not contemplate.

These sampling techniques allowed us to explore different subsets of the dataset and evaluate the performance of the models and NLP techniques under various sampling scenarios.

3.4.1 Text Classification – Sentiment Analysis

Firstly, based on the results reported by Table 3.6, we used MLMs (see Figure 3.4, *Language Models, Masked*). We chose XLM-T for experiments on Sentiment Analysis. XLM-T is based on XLM-RoBERTa (Conneau et al., 2019), further pre-trained on a corpus of 198 million tweets for 15 languages (Barbieri et al., 2022). This version of the model is specifically designed to handle the unique characteristics of tweets and social media posts, such as their limited length, informal language, and the presence of emojis and hashtags. More specifically, we use two variants of the XLM-T model:

- The base version (Barbieri et al., 2022) (XLM-T).
- XLM-T previously fine-tuned specifically for Sentiment Analysis (Barbieri et al., 2022) (XLM-T Sentiment). This sentiment variant has been already fine-tuned using 24,264 out-of-domain tweets in eight different languages, including French, English, and Spanish. However, it is important to note that these tweets cover a wide range of topics that do not necessarily include tourism.

We also experiment with the following generative LLMs (see Figure 3.4), *Language Models, Generative*):

- GPT 3.5 (Brown et al., 2020) (gpt-3.5-turbo-0125)[5], which is the latest version of GPT 3.5 with improved instruction following. This model is paid and closed source and was used through the OpenAI API[6]. Additionally, we also use GPT 4 (gpt-4-0125-preview) but only in zero-shot

[5] https://platform.openai.com/docs/models/gpt-3-5-turbo
[6] https://openai.com/blog/openai-api

settings due to API cost limitations. For both models, we use the default temperature of 1 and a presence penalty of 0.

- Mistral 7B (Jiang et al., 2023): We use Mistral-7B-Instruct-v0.2, the 7 billion parameters instruct version of the model. A batch size of 1, learning rate of 2e-4, and weight decay of 0.001 are used, as recommended in Massaron (2024b).

- LLaMA 2 7B (Touvron et al., 2023b): Similar to Mistral, we experiment with the 7 billion parameters instruct version of the model, namely, LLaMA-2-7b-chat-hf. The same hyperparameters as for Mistral are used, which were also recommended in Massaron (2024a).

LLMs (e.g., Mistral 7B, LLaMA 2 7B, GPT 3.5, and GPT 4), are applied in zero-shot and few-shot settings (FS), prompting the model with a few examples. We use the Sentiment Analysis prompt presented in Subsubsection 3.2.4. In the case of Mistral 7B and LLaMA 2 7B, as those models are open-source, we also experiment with fine-tuning them for Sentiment Analysis. We use the techniques and hyperparameter settings introduced in previous similar works (Massaron, 2024b,a).

With respect to the two MLMs, (e.g., XLM-T and XLM-T Sentiment), we used them as the backbone for three different training techniques (see Figure 3.4, *Machine-Learning Techniques*) along various dataset sizes using the sampling techniques described previously.

- *Fine-Tuning* (FT): optimal hyperparameters were found through grid search. For XLM-T 8 batch size, 2e-5 learning rate, weight decay 0.01; for XLM-T Sentiment 32 batch, 1e-5 learning rate and decay: 0.1.

- *Pattern-Exploiting Training* (PET) (Schick and Schütze, 2021) is used with default hyperparameters.

- *SetFit* (SF) (Tunstall et al., 2022): a prompt-free framework for few-shot fine-tuning of *Sentence Transformers*. We use the training parameters recommended in the SetFit repository[7], namely 4 epochs of training with a batch size of 16.

By comparing these deep learning-based techniques and assessing their effectiveness with varying amounts of annotated data, we aim to learn any insights about the minimum data requirements for achieving reliable Sentiment Analysis results in the tourism domain. This would allow us to establish which technique requires less annotated data to obtain competitive performance.

3.4.2 Sequence Labeling – Locations and Fine-grained Thematic Concept Extraction

For the sequence labeling task of NER for Locations, we experiment with and evaluate the following techniques.

- *Zero- and Few-Shot Sequence Labeling with Generative LLMs* (FS): we use the same generative LLMs as for Sentiment Analysis to have a common reference point, namely GPT 3.5, GPT 4, Mistral 7B, and LLaMA 2 7B.

[7] https://github.com/huggingface/setfit

- *EntLM* (Ma et al., 2022) with a multilingual BERT (mBERT) (Devlin et al., 2019) as backbone. The hyperparameters used are those recommended in the EntLM repository, namely a batch size of 4, learning rate of 1e-4, and weight decay of 0.

- For *Fine-Tuning* (FT) with MLMs, in addition to XLM-T, we include two additional models: XLM-RoBERTa (Conneau et al., 2019) (XLM-R) and the previously mentioned mBERT. Grid search for hyperparameter tuning found optimal values 8 for batch size, 5e-5 for learning rate, and 0.1 for weight decay. To fine-tune generative LLMs for sequence labeling we leverage the library published by García-Ferrero et al. (García-Ferrero et al., 2024). This fine-tuning technique allows performing sequence labeling tasks as a text-to-text generation task. We experiment with two generative models: LLaMA 2 7B and FlanT5 (more specifically flan-t5-base).

- *GoLLIE* (*Guideline following Large Language Model for Information Extraction*): a specialized language model trained to follow annotation guidelines (Sainz et al., 2024). It allows the user to perform sequence labeling inferences based on annotation schemes. The GoLLIE architecture can be enriched using domain-specific training examples. We will leverage the GoLLIE model with its base configuration of 7 billion parameters, pairing it with our training dataset.

In our analysis of Fine-grained Thematic Concept Extraction, we focused on two primary approaches: (1) employing the EntLM framework and (2) the Fine-Tuning (FT) of MLMs. Alternative approaches such as *Few-Shot Prompting with generative LLMs* and GoLLIE proved impractical due to the extensive number of thematic classes (e.g., 315 classes) involved. The sheer volume of classes exceeded the context window capacity of current models, leading to errors or the generation of random text unrelated to the task.

Moreover, we are interested in comparing the results of these approaches with the baseline established by the *word-matching* (rule-based) algorithm, as presented in Table 3.7. The primary objective of this comparison is to ascertain the minimum amount of annotated data required to justify transitioning from rigid *word-matching* approaches to more advanced deep learning techniques for each task, respectively. More specifically, we seek to determine the tipping point at which the benefits of employing deep learning techniques outweigh their data requirements. This is particularly true for a highly complex task such as Fine-grained Thematic Concept Extraction that involves labeling 315 different concept classes, and for which developing rule-based algorithms or manually annotating data are highly inefficient and expensive approaches.

Experiments were conducted on servers equipped with Nvidia A100 (80 GB VRAM) GPUs, Intel Xeon Gold 6226R CPUs (2.90 GHz), and 256 GB of RAM, and language models were accessed from the *HuggingFace* Transformers API (Wolf et al., 2020).

As it is customary, for Sentiment Analysis, we report accuracy results, while for sequence labeling we use the usual F1-micro metric calculated at the span level as defined in the CoNLL 2002 shared task (T. K. Sang, 2002). All reported results are the average of three randomly initialized runs.

3.5 Results

We will now report the results for the three tasks, namely Sentiment Analysis (Subsection 3.5.1), Named Entity Recognition (NER) for Locations (Subsection 3.5.2), and Fine-grained Thematic Concept Extraction (Subsection 3.5.3).

3.5.1 Sentiment Analysis

The results of the Sentiment Analysis on the five techniques are reported in Table 3.9. As a reminder, this task consists of classifying the polarity of each tweet as *positive*, *negative*, or *neutral*. We have highlighted in **bold** the results we will refer to in the text.

The most noteworthy aspect from the results is that fine-tuning XLM-T Sentiment (Table 3.9, *Fine-Tune of MLMs*) clearly outperforms any other method using only 5 examples for training (Table 3.9, 0.919). In contrast to previous work (Schick and Schütze, 2021), this suggests that fine-tuning on a large multilingual dataset for Sentiment Analysis, even with texts from different domains, dramatically helps improve the results in domain-specific touristic data, clearly outperforming few-shot prompting techniques with MLMs such as PET or LLMs. In fact, the fine-tuned XLM-T Sentiment model reaches optimal results with as few as 10 examples (Table 3.9, 0.939).

	Examples per class (positive, negative and neutral) — **Accuracy**								
Techniques	**0**	**5**	**10**	**20**	**30**	**40**	**50**	**100**	**All**
Prompt-based FS	Regular **Prompt-based Few-Shot** of LLMs								
GPT 3.5	**0.785**	0.739	0.757	0.766	0.694	0.685	0.664	0.645	
Mistral 7B	**0.716**	0.766	0.764	0.754	0.761	0.760	0.758	0.760	
LLaMA 2 7B	0.442	0.589	0.598	0.680	*Exceeding Input Context Length*				
FT of MLMs	Fine-Tune of **Encoder-Only Models** (MLMs)								
XLM-T		0.428	0.385	0.503	0.545	0.622	**0.646**	0.792	0.868
XLM-T Sentiment		**0.917**	**0.939**	0.922	0.877	0.875	0.925	0.914	**0.919**
FT of LLMs	Fine-Tune of **Encoder-Decoder and Decoder-Only Models** (LLMs)								
Mistral 7B		0.640	0.618	0.628	0.706	0.750	0.706	0.771	0.828
LLaMA 2 7B		0.594	0.651	0.613	0.738	0.763	0.759	**0.761**	0.844
PET	**Cloze-Style Few-Shot** with MLMs								
XLM-T		0.533	0.607	0.661	0.691	0.722	**0.764**	0.796	0.880
XLM-T Sentiment		0.598	0.717	0.729	0.819	0.787	0.855	0.874	0.877
SetFit (SF)	Combination of **Few-Shot and Fine-Tuning** for Sentence Transformers								
XLM-T		0.534	0.582	0.712	0.715	0.776	0.732	0.803	0.832
XLM-T Sentiment		0.831	0.878	0.876	0.893	0.882	0.899	0.858	0.821

Table 3.9: Sentiment Analysis with k-shot Sampling - Results on Text Classification Techniques (results in **bold** are referenced in the text)

Among the techniques that use only our domain-specific training data, Mistral 7B obtains the best scores with only 5 examples (PET with XLM-T requires 50 examples to obtain a similar score while SetFit performs similarly with 40-shot).

Among the methods using MLMs with only some examples from the training data, SetFit consistently outperforms fine-tuning until we reach 100 examples (e.g., Table 3.9), but with more data results from PET and fine-tuning are eventually better. Still, this highlights the effectiveness of SetFit, which is able to achieve high accuracy with only 40 examples and without requiring as much computing resources as for few-shot prompting with LLMs.

Looking at the performance of LLMs (Table 3.9, *Prompt-based Few-Shot*), we can see that in the zero-shot setting, GPT 3.5 and Mistral 7B obtain competitive accuracy ranging from 0.716 for Mistral 7B to 0.785 for GPT 3.5. We also experimented with GPT 4 in zero-shot settings. Interestingly, GPT 3.5 performed better than GPT 4 in zero-shot settings (0.785 for GPT 3.5 vs. 0.757 for GPT 4). We did not experiment with a higher number of shots with GPT 4 due to the high cost of the API, which is why it is not displayed in Table 3.9.

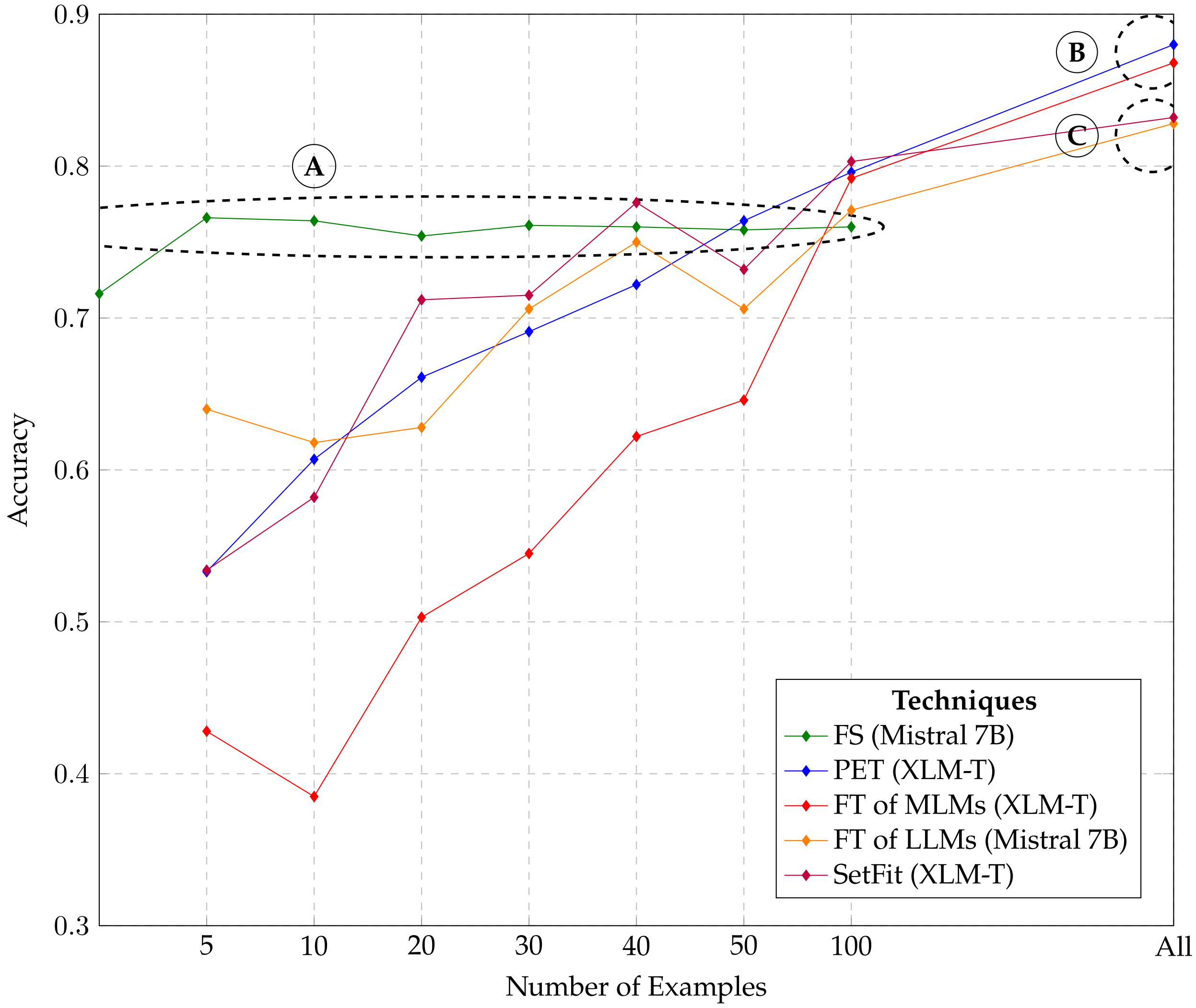


Figure 3.5: Sentiment Analysis - *k*-shot Sampling

Figure 3.5 provides another perspective on the results. Here, the *x-axis* represents the number of training examples used (in this case, tweets), while the *y-axis* indicates the accuracy scores. As we have observed previously, XLM-T Sentiment outperforms other models in most configurations because it has prior knowledge of the task. In Figure 3.5, we focus on the other models (namely, XLM-T and Mistral 7B) to analyze what would be the best technique in a text classification task where there is no available model with prior training like XLM-T Sentiment.

The observed patterns in Figure 3.5 and Table 3.9 point out the following four conclusions about Sentiment Analysis in the Tourism Domain:

1. When using an already fine-tuned sentiment model such as XLM-T Sentiment (see Table 3.9, *Fine-Tune with MLMs* and XLM-T Sentiment), a dataset containing as few as 10 examples is sufficient to achieve state-of-the-art Sentiment Analysis performance in the tourism domain.

2. When employing a base MLM such as XLM-T, SetFit appears to be a preferable choice in the tourism domain for few-shot scenarios, given that close to optimal performance can be reached with 40 examples per class (refer to Figure 3.5, 40 examples).

3. In use cases where very few annotated data is available (e.g., less than 30 examples per class) and no language model fine-tuned for the task exists (like XLM-T Sentiment), few-shot with LLMs such as Mistral 7B is the best option. This approach can produce an accuracy of roughly 0.750 consistently from 5 to 100 examples (refer to Figure 3.5, Ⓐ). It also produces robust results in zero-shot settings (accuracy of 0.716).

4. When no language model fine-tuned for the task exists, but a sizeable amount of annotated data is available, Pattern-Exploiting Training and Fine-Tuning of MLMs (refer to Figure 3.5, Ⓑ) produce slightly better results in the tourism domain than SetFit and Fine-Tuning of LLMs (refer to Figure 3.5, Ⓒ). Additionally, in contexts where annotated data are widely available, MLMs still produce the best results.

We will discuss the broader impact of these results later in Section 3.6. In the next sections, we present the results obtained for the sequence labeling tasks, namely NER for Locations and Fine-grained Thematic Concept Extraction.

3.5.2 Named Entity Recognition (NER) for Locations

Table 3.10 reports the results of NER for Locations. When the full training dataset is employed, all techniques yield comparable F1-scores (refer to Table 3.10, *All Examples*).

GoLLIE obtains the best overall result with a 0.832 in F1-score using the full training data. In zero-shot GoLLIE, Mistral 7B and GPT 3.5 perform quite similarly, between 0.670 and 0.694 in F1-score. While GPT 3.5 few-shot scores are slightly higher, Mistral is an open-source free model. Thus, annotating only 20 to 30 examples should be enough to obtain competitive results with Mistral 7B.

Fine-Tuning of MLMs with the full dataset produced F1-scores ranging from 0.791 with XLM-R to 0.818 with mBERT, and 0.808 with XLM-T (see Table 3.10). Notably, mBERT slightly outperforms the XLM model series in this task. However, fine-tuning requires a substantial volume of labeled data, as evidenced by low F1-scores when it has seen only a few examples.

	Examples per class (location) — **F1-score**								
Techniques	**0**	**5**	**10**	**20**	**30**	**40**	**50**	**100**	**All**
Prompt-based FS	Regular **Prompt-based Few-Shot** of LLMs								
GPT 3.5	**0.694**	0.698	0.762	0.762	0.798	0.809	0.828	0.806	
Mistral 7B	**0.680**	0.704	0.689	0.730	**0.749**	0.741	0.742	**0.739**	
LLaMA 2 7B	0.627	**0.587**	0.615	0.594	0.621	0.580	0.568	0.169	
FT of MLMs	Fine-Tune of **Encoder-Only Models** (MLMs)								
XLM-T		0.067	0.113	0.001	0.029	0.000	0.067	0.054	**0.802**
XLM-R		0.107	0.067	0.130	0.062	0.328	0.133	0.001	**0.791**
mBERT		0.115	0.108	0.083	0.007	0.000	0.000	0.000	**0.818**
FT of LLMs	Fine-Tune of **Encoder-Decoder and Decoder-Only Models** (LLMs)								
LLaMA 2 7B		0.000	0.000	0.000	0.000	0.000	0.000	0.228	**0.701**
FlanT5		0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.806**
EntLM	**Template-Free Few-Shot** in **Sequence Labeling** Tasks for MLMs								
mBERT		0.317	0.385	0.437	0.529	0.562	0.591	**0.584**	0.788
GoLLIE	**Guideline following model** for Information Extraction								
GoLLIE 7B	0.670	**0.622**	0.632	0.662	0.661	0.694	0.689	0.732	**0.832**

Table 3.10: Named Entity Recognition (NER) for Locations with k-shot Sampling - Results on Sequence Labeling Techniques (results in **bold** are referenced in the text)

Figure 3.6 provides a chart view of the k-shot sampling results. For each technique in Table 3.10, we report in the chart the results obtained with the most efficient open-source language model.

Summarizing, several key takeaways can be drawn from Figure 3.6:

1. With few examples, few-shot prompting with LLMs, such as GPT 3.5, produces the best results by far. It matches the word-matching approach (*reference F1-score*) in zero-shot settings and reach it with as few as 5 shots. Therefore, it should be prioritized when working with limited examples in the tourism domain. LLMs, out of the box, possess substantial knowledge about locations, likely due to their common exposure to similar tasks during training, which they can easily adapt. A few examples are sufficient to instruct these models on how to adapt this generic concept to domain-specific datasets. Alternatively, GoLLIE also achieves commendable results but only begins to surpass the rule-based approaches when provided with more than 50 examples per class. When using the full dataset, GoLLIE emerges as the optimal technique, achieving the highest performance, and thus should be favored with a larger volume of examples.

2. Both fine-tuning techniques perform poorly in contexts with a low number of examples, as indicated by their low F1-scores. Traditionally, fine-tuning with a large dataset has been the preferred technique for NER for Locations, but the advent of generative language-based few-shot prompting is beginning to shift this paradigm.

3. EntLM with MLMs yields more modest results but remains viable in 50-shot settings, unlike fine-tuning techniques.

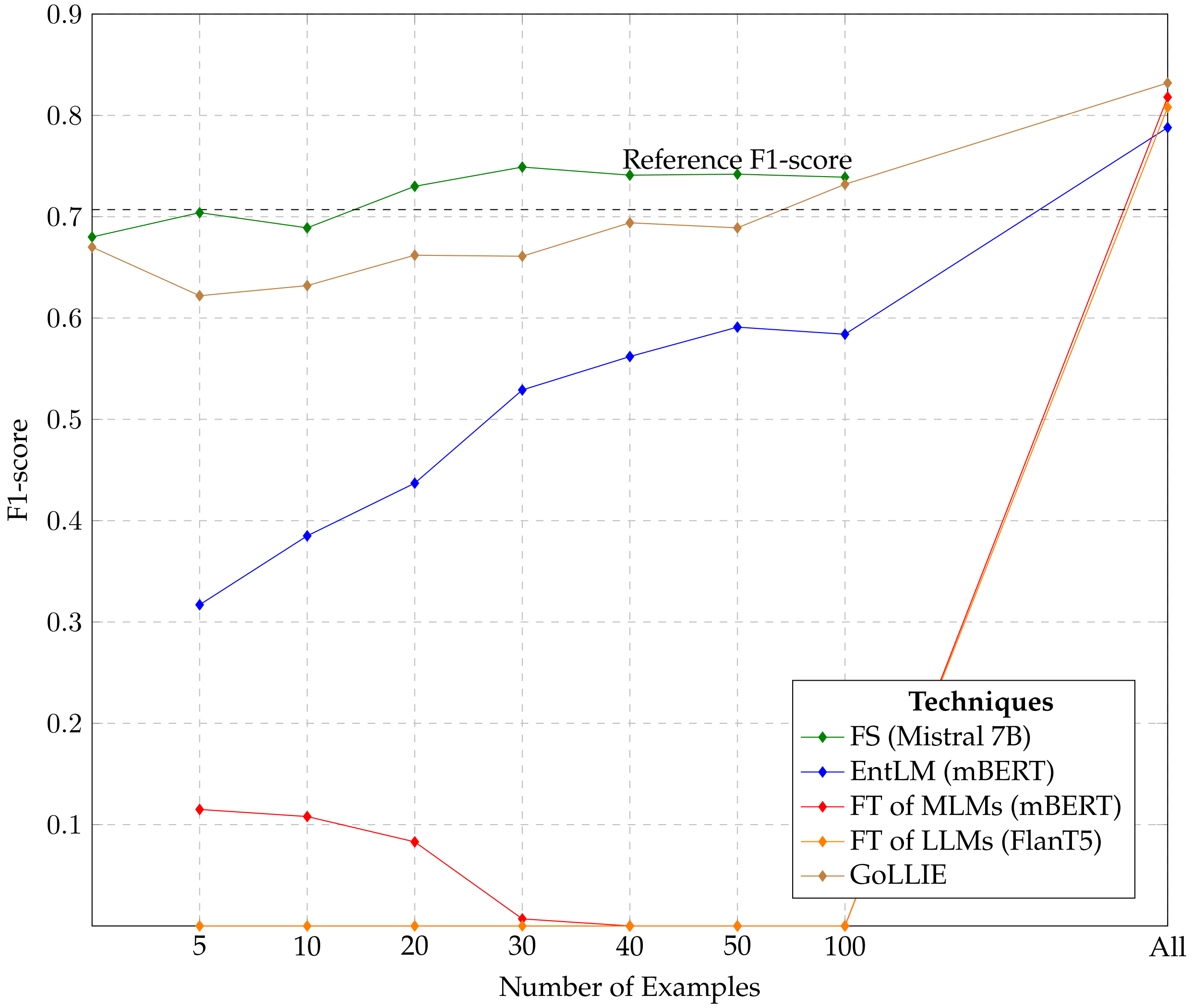


Figure 3.6: Named Entity Recognition (NER) for Locations – k-shot Sampling

Figure 3.7 shows selected techniques used with percentage sampling (LLMs did not fit due to context length constraints). In contrast to *k*-shot, it appears that percentage sampling is a better technique to set up EntLM and FT methods for sequence labeling. Thus, while EntLM is better with low amounts of data (5-10%), the fine-tuned models exhibit a similar upward trend with as little as 10% of the tweets, ending up outperforming EntLM as the number of data increases. We believe that the way this task is set is perhaps not the best fit for EntLM as the objective is to classify only one class, *location*, and EntLM's approach is based on generating *label words* which are associated with each entity class for better learning in few-shot settings.

However, this means that with only one class as a target, all the label words are assigned to the same class, generating a noisy signal which ultimately, as the number of words increases, hinders EntLM's performance.

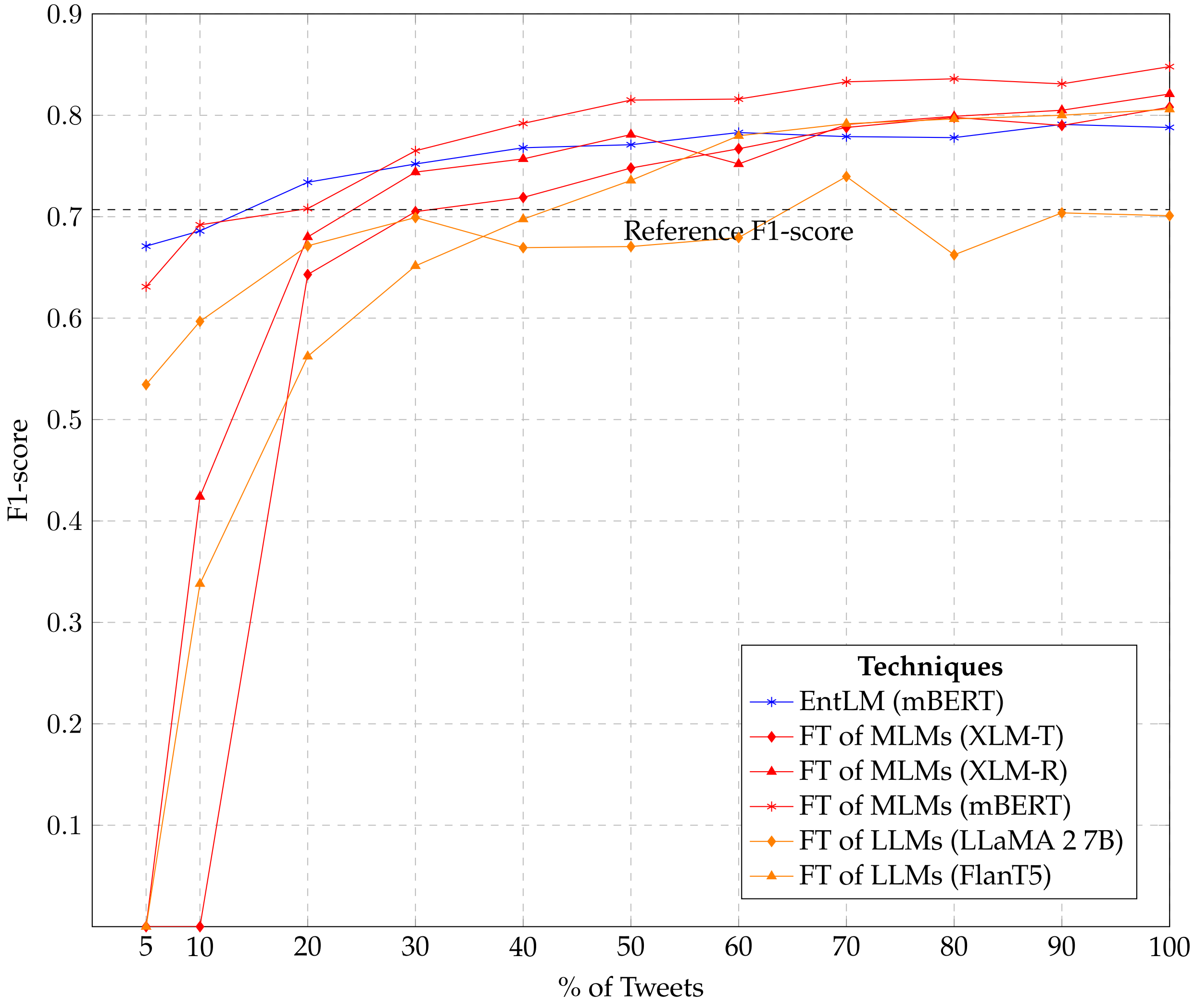


Figure 3.7: Named Entity Recognition (NER) for Locations – Percentage Sampling

As depicted in Figure 3.7, approximately 20% of the dataset (equivalent to about 330 tweets in our study) is necessary to achieve competitive results (e.g., outperforming the *word-matching* algorithm). Marginally inferior outcomes are observed with Fine-Tuning of LLMs, where LLaMA 2 7B slightly lags behind other models with an F1-score of 0.701, while FlanT5 aligns with the MLM fine-tuning results, achieving an F1-score of 0.806 with 100% of the training dataset used.

The EntLM technique shows that reliable results can be attained with less labeled data. Specifically, the fine-tuned mBERT does not surpass the performance of EntLM until more than 30% of the training data is used (as shown in Figure 3.7, 30%). Generally, fine-tuning becomes competitive only with percentage sampling, possibly due to the reduced frequency of O tokens in k-shot sampling. Conversely, EntLM demonstrates better performance with the same limited number of examples (as shown in Table 3.10, 0.584 with 100 examples).

Nevertheless, in scenarios with limited training examples, the performance of EntLM is not optimal. In such cases, our results indicate that prompt-based few-shot prompting techniques with LLMs excel (refer to Table 3.10, *Prompt-based Few-Shot*). Particularly in zero-shot settings, the best results are achieved with GPT 4 (note that it is not shown in Table 3.10, as we experimented with it only in zero-shot settings), which produces a remarkable F1-score of 0.829, equaling or surpassing all other strategies, even those involving extensively annotated datasets. Additionally, open-source (e.g., Mistral 7B, LLaMA 2 7B) or more cost-effective alternatives (e.g., GPT 3.5) also deliver satisfactory results in zero or few-shot settings (e.g., 0.694 for GPT 3.5, 0.680 for Mistral 7B in zero-shot settings), although they do not outperform fine-tuning techniques using the complete dataset.

Summarizing, optimal results in this task are achieved using Mistral with only 30 to 50 examples or, if computing requirements are too costly, with EntLM trained on 20% of the data, which amounts to 300 tweets. Having presented these findings, we now turn to the task of Fine-grained Thematic Concept Extraction.

3.5.3 Fine-grained Thematic Concept Extraction

Perhaps it is in the evaluation of Fine-grained Thematic Concept Extraction, shown in Figure 3.8 and Figure 3.9, where few-shot prompting with MLMs for sequence labeling clearly makes its mark. For a task that involves detecting and classifying sequences into a predetermined inventory of 315 classes, EntLM paired with mBERT performs very competitively.

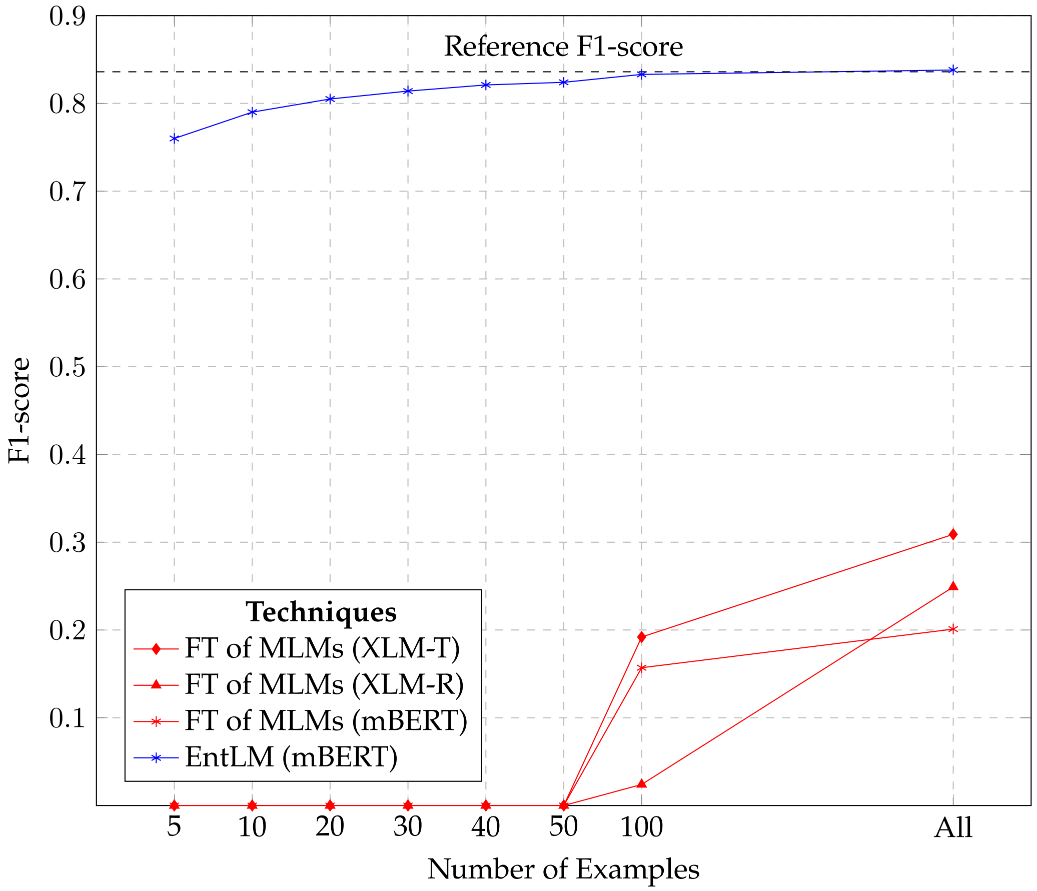


Figure 3.8: Fine-grained Thematic Concept Extraction – k-shot Sampling

Thus, with just five examples per class (5-shot setting), it obtains a 0.760 F1-score, almost equaling the *word-matching* algorithm's results with just a 50-shot training. These scores indicate a strong ability to accurately identify touristic concepts, as reflected by the high precision values spanning from 0.80 to 0.91.

Although overall results with *word-matching* were similar, EntLM was slightly superior in terms of recall while being slightly worse in precision. Still, EntLM's performance shows great promise to avoid costly manual annotation efforts or complex development of rule-based algorithms for domain-specific fine-grained sequence labeling tasks. EntLM's results are perhaps magnified by the very poor results obtained by the fine-tuning techniques in both data sampling scenarios, which indicates the difficulty of learning good sequence taggers for fine-grained tasks.

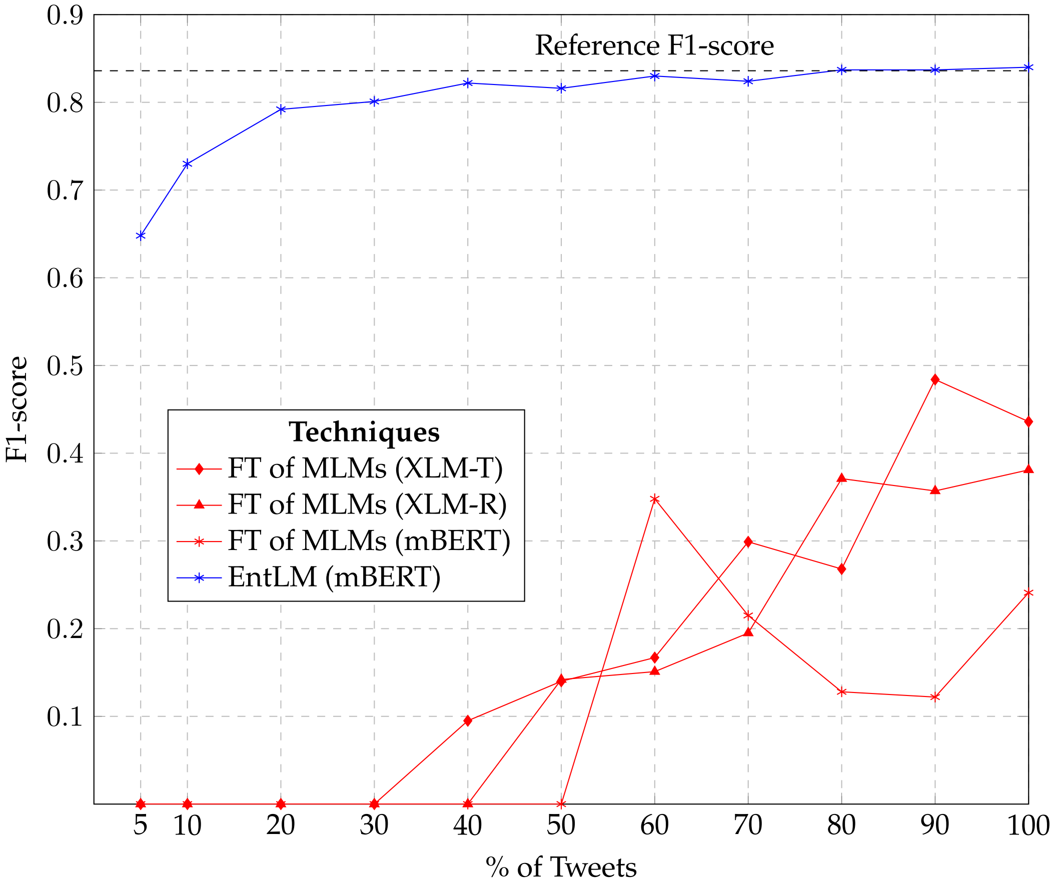


Figure 3.9: Fine-grained Thematic Concept Extraction – Percentage Sampling

Having conducted these experiments, we will now delve into the insights gained and discuss the key findings that emerged. Additionally, we will address the potential limitations and biases that might have affected the results.

3.6 Discussion and Limitations

In our comparative analysis of few-shot prompting and fine-tuning for three knowledge extraction tasks in the tourism domain, namely Sentiment Analysis (Subsection 3.6.1), Named Entity Recognition (NER) for Locations (Subsection 3.6.2), and Fine-grained Thematic Concept Extraction (Subsection 3.6.3), we have gained valuable metrics into the data requirements and performance of these techniques. However, our experiment also has some limitations (Subsection 3.6.4).

3.6.1 Sentiment Analysis

For Sentiment Analysis, our findings suggest that a model previously fine-tuned on a large out-of-domain dataset for the same downstream text classification task can outperform prompt-based techniques even in few-shot settings. If such extra data is not available for our target task, then addressing the task by means of few-shot prompting techniques with LLMs (in particular the GPT series of models and Mistral 7B) has demonstrated to be the best technique to avoid costly manual annotation work.

3.6.2 Named Entity Recognition (NER) for Locations

Regarding NER for Locations, the optimal strategy would be to use Mistral in few-shot settings with only 30 examples. Failing that, EntLM performs well when trained on percentage sampling using 20% of the training data.

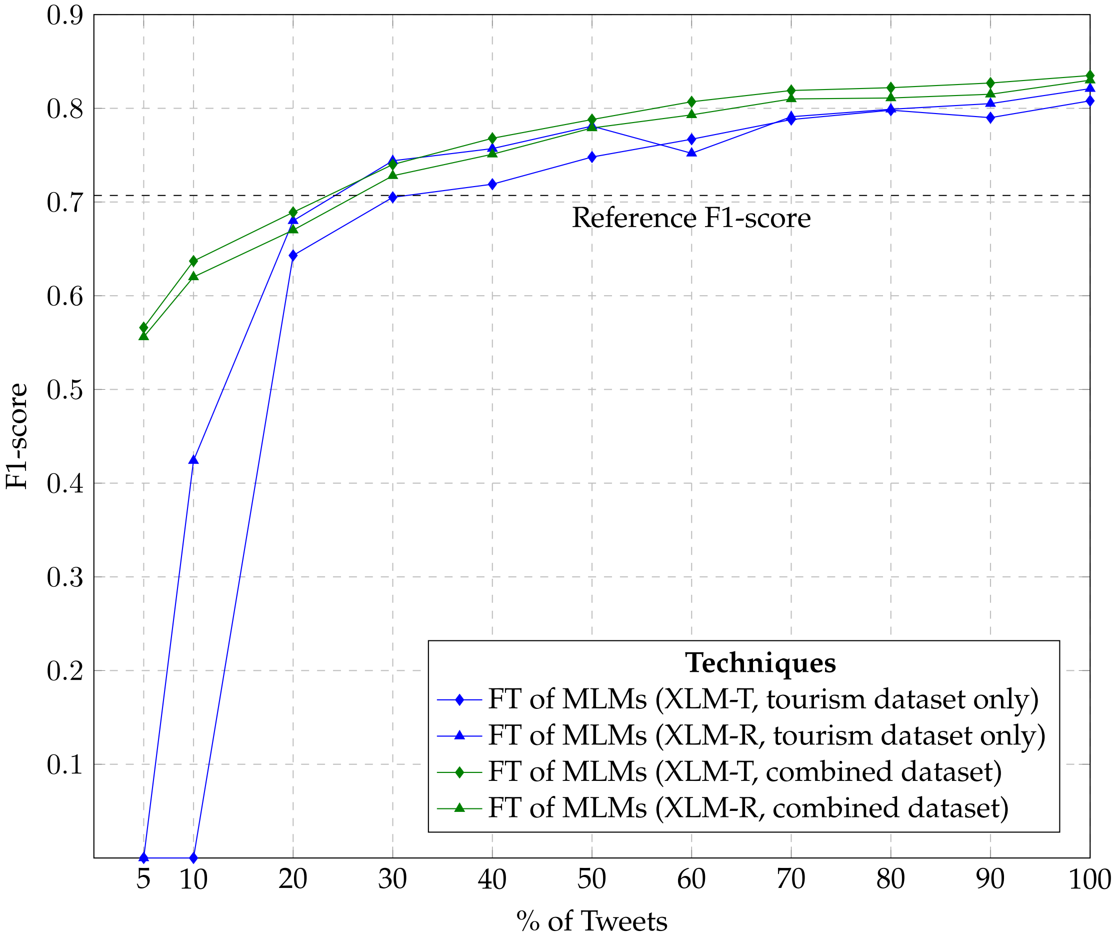


Figure 3.10: Comparison of Named Entity Recognition (NER) Model Performance: Our Dataset (*blue*) Compared to the Combined Dataset (*green*) – Fine-Tuning

The case of EntLM is interesting because in this task there is only a single class but with many label words associated with it (995 different location names). We hypothesized associating many label words to the same class does not benefit EntLM.

Interestingly, prompt-based few-shot prompting with LLMs and GoLLIE generally performs exceptionally well in contexts with limited examples. This effectiveness is largely due to their design, which leverages extensive pre-training on diverse data, likely containing a lot of location entities, allowing them to generalize this task from minimal input. Both the GPT series of models and Mistral 7B, for instance, have demonstrated very good results in these scenarios. Their ability to adapt quickly with little to no additional training data makes them the best techniques for zero-shot or few-shot settings.

We also explored the idea of improving our NER dataset by combining it with other existing NER (Named Entity Recognition) corpora from other domains, such as the Broad Twitter Corpus (Derczynski et al., 2016) (*English*), AnCora (Taulé et al., 2008) (*Spanish*) and ESTER (Galliano et al., 2006) (*French*), both already annotated with location entities. However, this experiment did not lead to any significant improvements in the F1-score (see Figure 3.10). In the figure, our dataset is highlighted in *blue* and compared to the combined dataset (incorporating ESTER, AnCora, and the Broad Twitter Corpus with ours) highlighted in *green* for the fine-tuning technique.

Upon merging the three datasets, the F1-score saw only a minor increase, rising from 0.808 to 0.835 for XLM-T and from 0.821 to 0.830 for XLM-R in fine-tuning scenarios. The limited improvement could be attributed to the fact that these corpora are not specifically designed for social media. ESTER consists of radio broadcast transcripts, while AnCora comprises newspaper texts. Consequently, they lack the contextual information pertinent to the tourism domain.

3.6.3 Fine-grained Thematic Concept Extraction

In the case of Fine-grained Thematic Concept Extraction, it is a different and more complex sequence labeling task that involves a large inventory of classes (315 concepts instantiated out of the 1,494 from the WTO tourism thesaurus), each having very few instance label words that are highly representative of the classes to which they refer. Figure 3.11 shows the low count of unique label words for the fine-grained thematic concepts most often found in the tweets.

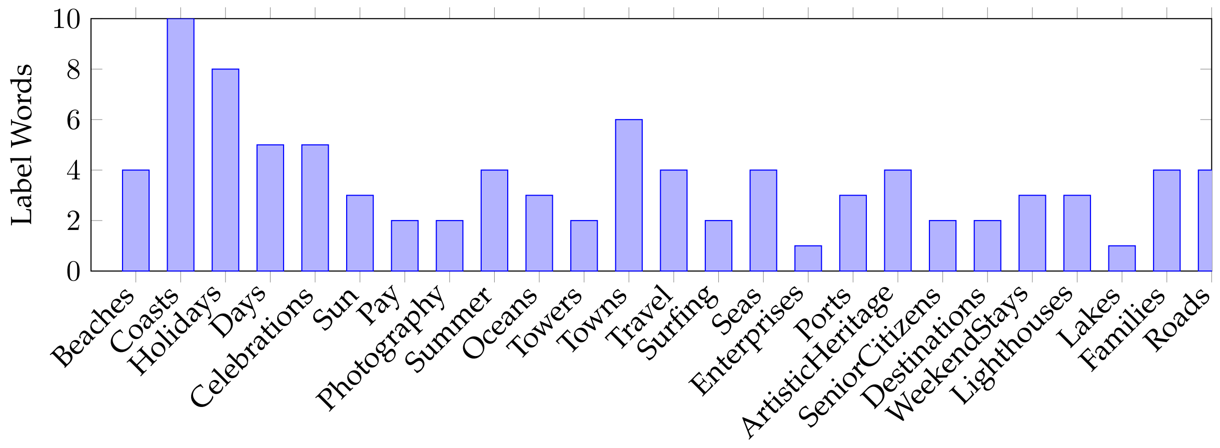


Figure 3.11: Number of Label Words for the Most Frequent Fine-grained Thematic Concepts Found in the Tweets

As we expected, fine-tuning did not yield any satisfactory results regardless of the amount of data used or the sampling technique employed. We believe this is due to the low number of examples per class. However, a major finding is that the EntLM few-shot prompting techniques demonstrated solid performance for this task, even when trained with a limited number of examples on a very small percentage of the available data. This robustness could be attributed to the model's ability to generalize effectively from a smaller set of label words which are associated to each of the classes, namely, the thematic concepts. This result is particularly useful for practical applications where manually annotated data is usually very scarce.

3.6.4 Limitations

Our experiments have some limitations. Firstly, we intentionally focused on three common knowledge extraction tasks within the tourism domain, ensuring an in-depth understanding of this area. Although our findings are specific to these tasks, additional studies can broaden the applicability to other domains. Second, we worked with a curated dataset of 2,961 tweets. This limitation in dataset size was purposefully selected to maintain focus, but larger and more diverse datasets might offer further perspectives. Lastly, while our dataset has a diverse language representation, it predominantly features French tweets, mirroring the demographics of visitors in the *French Basque Coast* region.

This study allowed us to determine the best techniques for each of the three knowledge extraction tasks (Sentiment Analysis, NER for Locations, and Fine-grained Thematic Concept Extraction) and the number of examples required to obtain optimal results in the tourism domain. We will use the generated annotations in the APs Framework to extract structured information from natively unstructured social media posts and instantiate our data model. This model and its applications will be presented in the next chapter (see Chapter 4). Before moving to this, let's conclude the *Transform* phase.

3.7 Summary and Perspectives

In this chapter, we explored the most effective strategies for processing social media content in the tourism domain. More precisely, we focused on three recurring tasks of knowledge extraction in this domain: Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction. Firstly, given the scarcity of existing annotated multilingual corpora in this domain, we have contributed a novel multilingual (French, English, and Spanish) social media dataset. This dataset is annotated at the text level for sentiment and at the token level for locations and touristic fine-grained thematic concepts, based on the *Thesaurus on Tourism and Leisure Activities of the World Tourism Organization* (Contribution 2.1). Then, using this dataset, we present a comparative study of various approaches (e.g., rule-based, fine-tuning, few-shot prompting) and language models (e.g., XLM-RoBERTa, mBERT) specific to the tourism domain (Contribution 2.2). The primary goal of this study is to identify the best strategy for achieving competitive results while minimizing the necessity for time-consuming and costly manual annotations, especially in complex knowledge extraction tasks such as Fine-grained Thematic Concept Extraction.

Experiments with various Masked Language Models (encoder-only MLMs) and Large Language Models (LLMs, decoder and encoder-decoder) in few-shot and fine-tuning settings demonstrate that it is possible to obtain competitive results for all three tasks with very little annotation data: 5 tweets per label (15 in total) for Sentiment Analysis, 30 examples using generative LLMs for NER and 1k tweets annotated with fine-grained thematic concepts. Our results also suggest that few-shot prompting for sequence labeling (Ma et al., 2022) using MLMs seems to be particularly effective for highly fine-grained tasks (more than 300 classes). Overall, the obtained results indicate that MLMs applied in few-shot settings remain competitive with respect to LLMs for discriminative tasks. This is coherent with previous results published for these types of tasks (Tunstall et al., 2022). We believe that these findings are beneficial not solely for the advancement of our application but also for other domain-specific applications that necessitate NLP analysis for model enrichment.

However, future research on similar NLP tasks in different domains and more languages is required to establish the transferability of our findings beyond the tourism domain and to other types of textual data (e.g., well-formatted and longer texts like newspapers). Such investigations would be valuable not only for NLP researchers specializing in tourism but also for other applications requiring domain-specific NLP analysis, particularly in scenarios where annotated data is scarce or there is a desire to move away from ad-hoc rule-based methods.

Additionally, it would be interesting to propose approaches to use few-shot prompting with LLMs for the task of fine-grained thematic concept extraction. The main issue right now is that these models lack the context window required to handle a large number of classes (in our case, 315 classes). We propose to investigate and compare two approaches in the future:

- *Hierarchical Approach*: Group the fine-grained thematic concepts into more coarse categories (e.g., all sea-related concepts are grouped together) to reduce the number of labels (and therefore the prompt). Run inference in two steps: first, with coarse thematic concepts, then with fine-grained thematic concepts separately for each coarse label.

- *Batching Approach*: Batch the labels arbitrarily to run inference on a smaller subset of labels at a time (e.g., 20 labels per inference) and then merge the results from all these batches.

We will now move on to the next phase of the APs Framework (*Analyze* in Figure 1.5) and explain how fine information extracted from posts can be used to instantiate a domain-independent model for social media and calculate indicators for stakeholders across various domains.

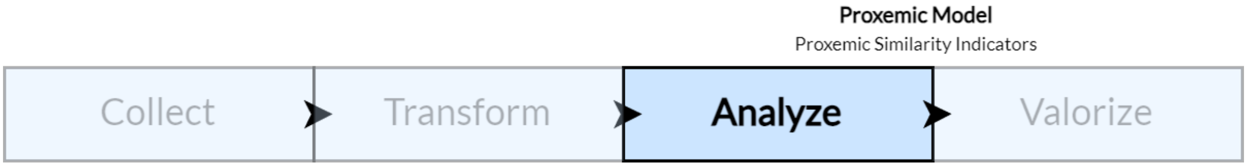
Chapter 4

Analyze

Redefining *Proxemics* to Model Social Media Entities and their Interactions to Generate Domain-Adaptable Indicators from Social Media

"Everything is related to everything else, but near things are more related than distant things."
— Waldo Tobler, Geographer

The broad diversity of available data on social media makes it valuable for analyzing and gaining knowledge into a wide array of vastly different domains (Rathore et al., 2017). This chapter, associated with the *Analyze* phase of the APs Framework (Figure 1.5), is positioned within the *Data Analytics* and *Decision Support Systems* fields and addresses the challenge of modeling social media interactions in a domain-agnostic manner to produce adaptable indicators applicable to various application domains.



Here, we hypothesize that adapting the *proxemics* theory (Hall, 1966; Hall et al., 1968) and proxemic dimensions (Greenberg et al., 2011) to social media could provide a generic and universally applicable way to model social media entities and their interactions to produce relevant domain-independent indicators.

We start this chapter with a brief introduction highlighting the growing requirement for social media indicators applicable to various domains (Section 4.1). We then review existing works calculating indicators on social media data for end-users (Subsection 4.2.1). We notice that most of them are heavily focused either on specific domains or on particular tasks, making them unsuitable for direct use in our generic framework. Stemming from this, and to alleviate this problem, we propose leveraging the *proxemics* theory (Hall, 1966; Hall et al., 1968) for this purpose. We review existing uses of the *proxemics* theory in various physical spaces (Subsection 4.2.2) and then propose a formal redefinition of this theory and associated dimensions (*Distance*, *Identity*, *Location*, *Movement*, and *Orientation*) for use within digital social media spaces (Section 4.3, Contribution 3.1), along with a proxemic data model: the APs Trajectory Model (Masson et al., 2023b) (Section 4.4, Contribution 3.2). But we are aiming to go further than simply modeling social media interactions and want to design domain-adaptable indicators based on this redefinition. To this purpose, we review existing similarity measures (Subsection 4.5.1) and criteria combination techniques (Subsection 4.5.2) to propose a toolkit, *ProxMetrics* (Masson et al., 2024b), and associated formula to express domain-adaptable social media indicators as composite proxemic similarity measures (Section 4.6, Contribution 3.3). Our proxemic toolkit is then experimented in the domain of tourism (Section 4.7); we instantiate the model using our dataset, and then, using requirements collected from a local tourism office, leverage the toolkit to design indicators in the domain of tourism (Subsection 4.7.1), we do a detailed case study on specific indicators (Subsection 4.7.2 and Subsection 4.7.3), and then evaluate qualitatively the results (Subsection 4.7.4). Lastly, we conclude by proposing perspectives to extend this work (Section 4.8). This novel application of *proxemics* was presented at the following international conference and journal:

- M. Masson, P. Roose, C. Sallaberry, R. Agerri, M. N. Bessagnet, A. Le Parc Lacayrelle. (2023). APs: A Proxemic Framework for Social Media Interactions Modeling and Analysis. In *International Symposium on Intelligent Data Analysis (IDA 2023)* (pp. 287-299) (Louvain-La-Neuve, Belgium). Cham: Springer Nature (CORE Rank: B, ERA Rank: A).

- M. Masson, P. Roose, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, R. Agerri. (2024). ProxMetrics: Modular Proxemic Similarity Toolkit to Generate Domain-Adaptable Indicators from Social Media. In *Social Network Analysis And Mining*, 14, 124. Springer (Impact Factor: 2.8).

4.1 Introduction: From Domain-Specific to Domain-Adaptable Indicators on Social Media

Social media have transformed from simple digital platforms for informal conversations to massive networks that shape modern society (Akram and Kumar, 2017). They now encapsulate a broad spectrum of daily life, ranging from communication and commerce to politics and entertainment, serving as a valuable data source that provides insights into how people perceive and engage with the world. The versatility of social media data makes it valuable for analyzing and gaining insights into a wide array of vastly different domains, including but not limited to *tourism*, *public policy*, and *healthcare*. Analysis of social media data often involves the calculation of indicators (Neiger et al., 2012), which are metrics that aim to measure or evaluate the state or level of a particular aspect of interest (e.g., the affluence associated with a given place (Khalifa et al., 2017), the level of friendship between users (Aiello et al., 2012), etc.). Typically, indicators are defined based on the domain of application.

- In the *domain of tourism*, stakeholders are increasingly calculating indicators based on social media data for various purposes (Hvass and Munar, 2012). These purposes range from analyzing frequently practiced leisure activities and their correlations with climatic or temporal factors, to understanding typical touristic routes and gauging levels of satisfaction. This data can assist in tasks such as the creation of tailored tour packages or the identification of touristic areas requiring refinement.

- In the *domain of public policies*, social media indicators help contextualize public sentiment, enabling government agencies to make informed decisions (Charalabidis and Loukis, 2012) to improve public services, urban planning, or address societal issues.
- In the *domain of healthcare*, social media provide a platform for patients to share their experiences and for healthcare providers to disseminate information and monitor health trends through indicators, ultimately contributing to a better understanding of medical interventions and patient outcomes (Smailhodzic et al., 2016).

Indicators are usually related to heterogeneous social media entities. These entities may be directly associated with the social media platform, such as individual users (Constantinides et al., 2010), groups of users (Esmaeili et al., 2011), posts, etc., but they can also be informational entities extracted from the content or metadata of social media posts, including themes (Yadav and Sagar, 2023), events (Aldhaheri and Lee, 2017), persons (Ma et al., 2019), medical entities (Scepanovic et al., 2020), organizations, etc.

A prevalent category of indicators is similarity measures. These are quantitative indicators used to assess the degree of *resemblance* or *closeness* between several entities. For instance, in the context of social media, a similarity measure might compare user profiles based on shared features (e.g., similar age, same gender, nearby home location, same language) to suggest potential connections with other users or content recommendations (Mazhari et al., 2015).

Similarity measures are crucial in various domains for many applications, including content recommendation (Jiang and Yang, 2017), targeted advertising (Zhang et al., 2019), understanding social dynamics, and event detection (Becker et al., 2010).

When interpreted correctly, they can provide insights into user connectivity and preferences, the nature of interactions, and common patterns found online. However, determining what *similarity* means in the context of social media can be challenging. Different platforms, users, and objectives lead to multifaceted interpretations of similarity (Anderson et al., 2012).

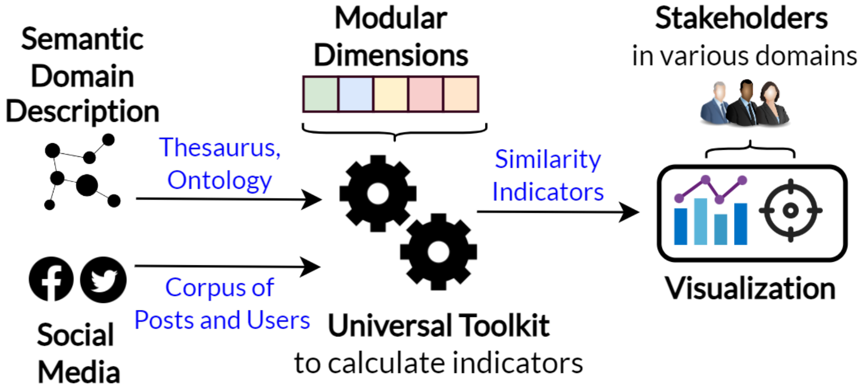

Semantic Domain Description
Thesaurus, Ontology
Modular Dimensions
Similarity Indicators
Stakeholders in various domains
Visualization
Social Media
Corpus of Posts and Users
Universal Toolkit to calculate indicators

Figure 4.1: The Requirement for Domain-Adaptable Indicators in the Context of the APs Framework

Figure 4.1 shows our requirements in the context of the APs Framework, where we have, as input, data feeds from social media and a semantic description of the domain of interest. We require a universal toolkit that can compute indicators adaptable to a variety of application domains, for example, through a composite approach with modular dimensions. The notion of "*indicators*" is very broad; therefore, in this work, we have decided to focus on similarity indicators. The produced similarity indicators would then be visualized for stakeholders in various domains.

We will now review existing work on calculating indicators on social media to determine if any versatile and domain-adaptable options exist, that we could leverage.

4.2 Related Work: Social Media Indicators and *Proxemics*

In this related work section, we begin by reviewing existing works that leverage indicators on social media across various domains (see Subsection 4.2.1). Given their heavy focus on single domains or very specific tasks, we then present and examine the use of the *proxemics* theory (Hall, 1966) to determine its potential as a foundation for calculating composite, domain-adaptable indicators from social media (see Subsection 4.2.2).

4.2.1 Review of Existing Social Media Indicators

Figure 4.2 is a mind map we have designed to present the primary categories of indicators employed in social media research, it is organized into six broad categories: (1) *User Behavior*, (2) *User Demographics*, (3) *Engagement Metrics*, (4) *Conversion Metrics*, (5) *Content Analysis*, and (6) *Platform Performance*. Indicators featuring a *red* border are those we chose to reuse in our work, more details will be given later. It is crucial to acknowledge that these indicators are not exclusively computed through automated data analytics processes. Many fields within the social sciences, such as psychology (Huang, 2017), also use social media indicators, often through manual analysis of posts.

Indicators on Behavior Indicators centered on user behavior (refer to Figure 4.2, ①) encompass analyses of user mentions by others on social media (Sadri et al., 2018), the profiles that certain users tend to visit more frequently or less so (Benevenuto et al., 2009), patterns of activity on social media, for example, when browsing from mobile devices (Priambodo and Satria, 2012), and the duration spent on these platforms. Furthermore, some research has delved into the nature of interactions on social media (Wilson et al., 2009), for example, between students (Yohanna, 2020).

Indicators on Demographics Certain indicators specifically target user demographics and characteristics (Cesare et al., 2017) (refer to Figure 4.2, ②). For instance, some studies focus on age-related features to examine the behavior of adolescents (Vannucci et al., 2020) or older users (Bell et al., 2013). Age-based indicators are particularly relevant in political studies for exploring political interests and participation (Holt et al., 2013). Gender has been extensively researched as well, primarily for developing marketing-related indicators (Hudders and De Jans, 2022), investigating gender identity on social media (Bamman et al., 2014), or analyzing the impact of gender on social media imagery (Rose et al., 2012). The language used by users represents another notable aspect, whether it be for studying the mental health of individuals through their social media activity (Gkotsis et al., 2016) or for automatic personality assessment (Park et al., 2015). Other areas of interest include determining the home geolocation for user profiling (Chen et al., 2016), identifying the interests of social media users (Kang and Lee, 2017), or understanding the devices primarily used (e.g., *desktop*, *mobile*) (Humphreys, 2013).

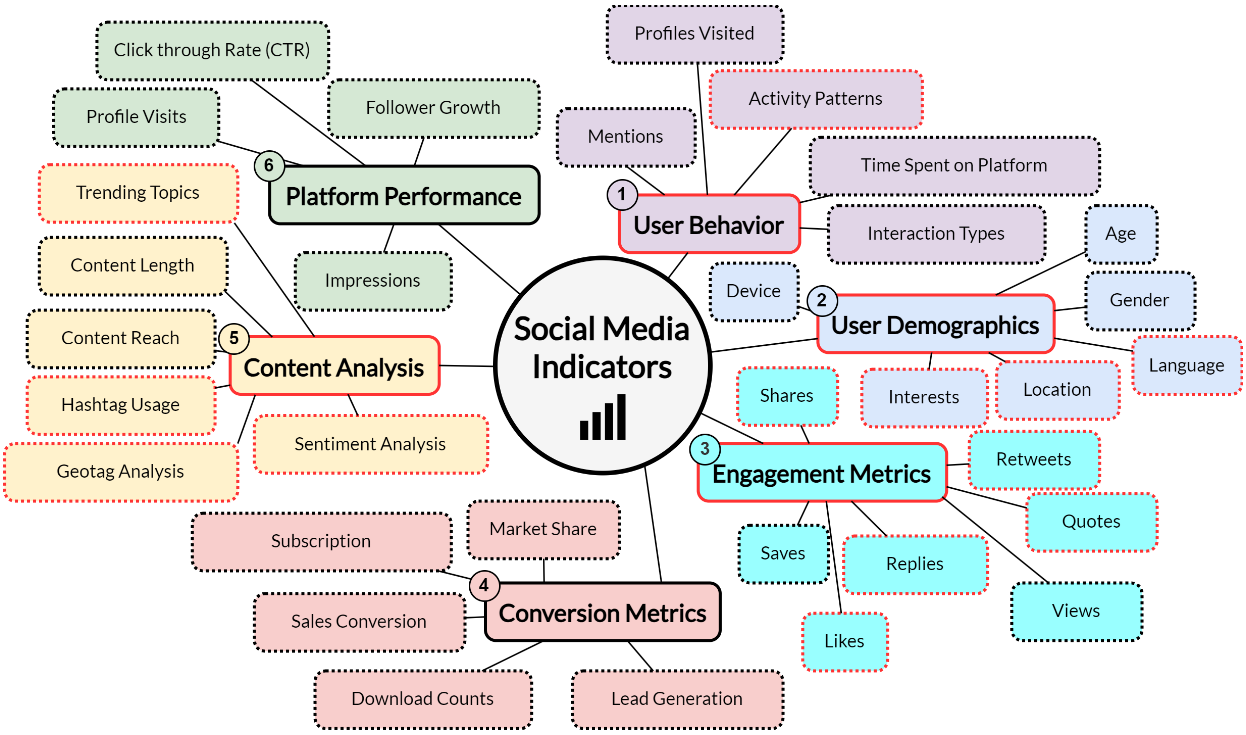


Figure 4.2: Mindmap of the Main Categories of Existing Social Media Indicators. Indicators featuring a *red* border are those we chose to reuse

Indicators on Engagement Another type of indicator is engagement metrics (refer to Figure 4.2, ③) (Trunfio and Rossi, 2021); these indicators are more straightforward to use. Engagement includes many features depending on the social media platform, like likes, retweets, replies, quotes, views, shares, etc. These metrics have been used in risk communication (Kim, 2021), to assess the strategies of specific types of social media accounts, like those related to food (Barklamb et al., 2020), or to study purchase intention (Rahman et al., 2017).

Indicators on Conversion Conversion metrics (refer to Figure 4.2, ④) are indicators used in digital marketing analytics to measure the effectiveness of marketing campaigns (Misirlis and Vlachopoulou, 2018), website performance, and overall business objectives in converting visitors into desired actions. They can be based on the number of subscriptions or downloads generated (for example, in the research domain, the number of paper downloads and citations (Tonia et al., 2016)), conversion into sales, progression in market share (Sherly et al., 2020), or lead generation (Evans et al., 2021) based on social media strategies.

Indicators on Content Indicators based on content analysis (refer to Figure 4.2, ⑤) are more predominant in computer science due to their complexity to perform manually. They can involve the analysis of trending topics and hashtags, such as disaster-related ones (Murzintcev and Cheng, 2017), the length and quality of content (Gkikas et al., 2022; Agichtein et al., 2008), or content reach (Parsons, 2013). Sentiment analysis is also widely used (Yue et al., 2019), for instance, to better understand people's feelings toward COVID-19 (Nemes and Kiss, 2021) and vaccine hesitancy (Piedrahita-Valdés et al., 2021), as well as geospatial analysis (Owuor and Hochmair, 2020), for example, using geotags metadata in posts (Schulz et al., 2013), like to assess park visitation and equitable park access (Hamstead et al., 2018).

Indicators of Platform Performance Finally, some indicators are based on platform performance (refer to Figure 4.2, ⑥). For example, through the number of profile visits, and content impressions, including studies on the reach of information on sickness to social media users (Theiss et al., 2016) and, more generally, to study the effect of advertisements on those platforms (Lee et al., 2018). Follower growth is analyzed to study humanitarian organizations' sharing of content (Yoo et al., 2020) or athletes' social media following base (Bredikhina et al., 2023), or click-through rate, including metrics such as *Clicks Per Follower* (CPF) and *Clicks Per Impression* (CPI) on links (Wang et al., 2016a).

Discussion: The Requirement for Composite, User Customizable Indicators

The landscape of social media indicators, as outlined in the preceding subsections, shows that a wide range of metrics is available to stakeholders, covering many requirements. However, several challenges emerge upon closer examination of these indicators, especially in the context of our domain-adaptable, generic framework (APs Framework):

1. *Domain Specialization:* Many of these indicators are applicable primarily in specific domains, such as conversion metrics, which are mainly used in financial and marketing domains. While this specialization is beneficial for targeted analyses, it inherently limits the cross-applicability of indicators, rendering them less effective for interdisciplinary research or applications spanning multiple domains. In the context of our framework, we require indicators that are easily domain-adaptable.

2. *Redundancy:* Another issue is the potential redundancy among these indicators. For example, in many domains, likes and shares represent similar metrics, as do hashtag usage and trending topics. There is no requirement to delve deeply into this level of detail. This complicates the interpretation of results for non-computer scientist users. A streamlined set of indicators bringing complementary insights would facilitate analyses for non-computer scientist users.

3. *Difficulty in Integrating Various Indicators for Complex Requirements:* The current set of indicators remains fragmented, posing challenges for users, particularly those with minimal computing expertise, in effectively combining them. Our framework requires the development of a mechanism that facilitates the creation of composite indicators (e.g., multi-criteria indicators that are constructed by combining various individual indicators into a single one), customizable to meet the unique requirements of different application domains.

We have selected base indicators to reuse, highlighted in red in Figure 4.2, based on several criteria. These indicators are (1) generic enough to be applicable across various contexts within our framework, (2) broadly used in social media analysis, and (3) computationally straightforward to automate (e.g., not based on complex, multi-criteria analyses), thus not necessitating manual content analysis or surveys (in the long term, we want our framework to be fully automated). This selection aims to incorporate metrics that are fundamental and widely recognized for their relevance and utility in social media analytics.

To address these challenges, we hypothesize that adapting the theory of *proxemics* for social media could provide a generic, domain-adaptable foundation to compute composite indicators on social media that could be integrated into our framework. We briefly introduced this theory in Chapter 1, but we will now delve further into what the *proxemics* theory entails.

4.2.2 The *Proxemics* Theory

Proxemics was introduced in the seminal work of the American anthropologist Edward T. Hall (Hall, 1966). He defines *proxemics* as:

> *The science that studies the organization of space and the effect of distance on interpersonal relations.*

Hall studied physical distance and the way it affects and regulates interactions between people. He then went further and linked the concept of distance to *proxemic zones* (Hall et al., 1968). There are four core proxemic zones (see Figure 4.3). It is crucial to note that cultural, social, and physical factors can affect the definition of these zones.

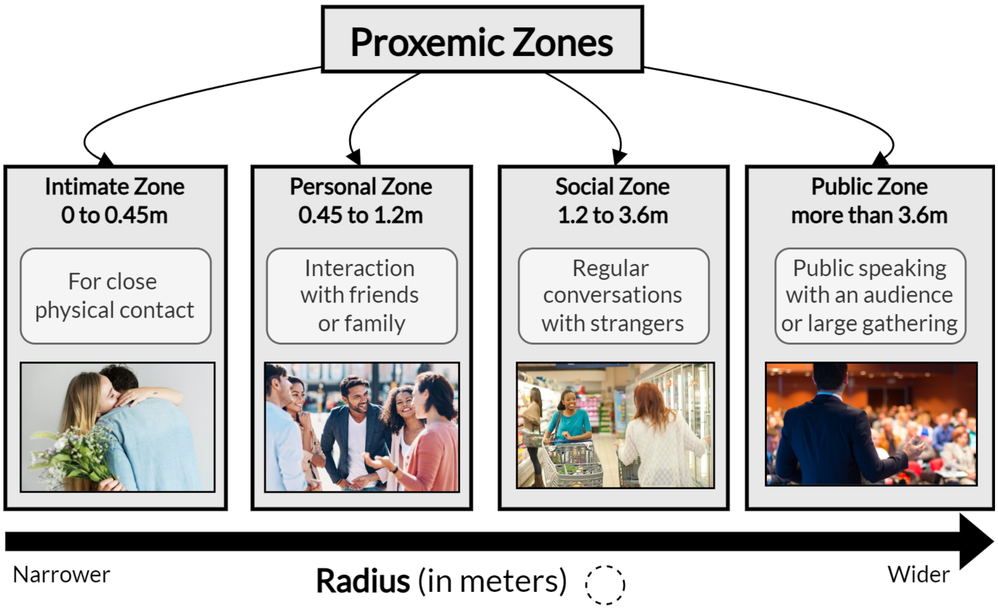

Proxemic Zones
Intimate Zone
0 to 0.45m
For close physical contact
Personal Zone
0.45 to 1.2m
Interaction with friends or family
Social Zone
1.2 to 3.6m
Regular conversations with strangers
Public Zone
more than 3.6m
Public speaking with an audience or large gathering
Narrower
Radius (in meters)
Wider

Figure 4.3: Edward T. Hall's Four Proxemic Zones

The Five Proxemic Dimensions (DILMO)

In 2011, Greenberg *et al.* extended Hall's definition of *proxemics* to introduce the notion of *proxemic dimensions* (Greenberg et al., 2011) (also referred to as the DILMO dimensions).

They identified five dimensions that can be used to express *proxemics* (Distance, Identity, Location, Movement, and Orientation). These dimensions are presented in Table 4.1.

Dimension	Definition	Example
Distance (D)	The measure of separation between several entities (e.g., *persons, objects*).	Physical distance in meters (*numerical*). Whether two entities are in the same room or not (*categorical*).
Identity (I)	A set of characteristics describing the individuality and the role of an entity.	Age, height (*numerical*). Name, gender (*categorical*).
Location (L)	A qualitative description of the space. Position of static (e.g., *furniture*) and dynamic (e.g., *persons*) entities.	Euclidean coordinates: x, y, z (*numerical*). Room in which an entity is (*categorical*)
Movement (M)	The change of location and orientation over time.	Spatio-temporal sequence (*numerical*). Descriptors of speed such as fast, slow, rapid, or steady (*categorical*).
Orientation (O)	The direction in which an entity is facing.	Bearing (*numerical*). Facing toward or away from something (*categorical*).

Table 4.1: The Five Dimensions of *Proxemics* (DILMO), as Defined by Greenberg et al. (2011)

Proxemic Analysis Levels and Centrality

Proxemics can be studied at several levels: (1) the individual level (*how and why does an individual express specific traits and cognitive or affective states through their proxemic behavior?*) and (2) the group level (*how does the behavior of individuals affect the group?*) (McCall, 2015).

A crucial concept in *proxemics* is known as *centrality*. *Proxemics* requires the selection of a reference, or central entity (e.g., a specific individual, an object, etc.) that will serve as the focal point for behavioral observations and analyses (Pérez et al., 2021). For example, in an urban planning context, a central landmark building can be chosen as the reference entity in a municipality. Urban planners might study the behavior and movement patterns of individuals in and around this central entity to gain insight into how people use the space, how they interact with each other, and how other entities (like nearby businesses) relate to it.

Existing Works Leveraging *Proxemics*

When it comes to usage, *proxemics* is used primarily to analyze physical interactions using tangible, physical metrics, as shown in Table 4.2. For example, for robot navigation (Rios-Martinez et al., 2015), classroom behavior analysis (Castañer et al., 2013) or to design proxemic-aware mobile applications linked with sensors (Pérez et al., 2021).

The concept has also been extended to other applications like picture annotation (Yang et al., 2012) and assessing the impact of social distancing during COVID-19 (Mehta, 2020). Recent works have tried adapting *proxemics* to virtual spaces but typically still relied on physical metrics, such as in video games or Virtual Reality (VR) worlds (Llobera et al., 2010). Recently, the concept of digital *proxemics* has emerged, focusing on non-physical interactions in virtual spaces, such as in the analysis of cybercrimes (Gunawan et al., 2021) or for middleware reconfiguration (Luxey, 2019).

Reference	Space	Metrics	Level	Application Domain
Llobera et al. (2010) *Detect people reaction in a VR world.*	Cyber	Physical	Individual	Virtual Reality
Cristani et al. (2011) *Social relation inference.*	Physical	Physical	Group	Psychology
Yang et al. (2012) *Picture annotation.*	Physical	Physical	Individual	Psychology
Castañer et al. (2013) *Study teachers' behaviors.*	Physical	Physical	Both	Education
Mueller et al. (2014) *Proxemic strategies for videogames.*	Physical	Physical	Individual	Video Games
Hans and Hans (2015) *Analysis of non-verbal communication.*	Physical	Physical	Individual	Psychology
Rios-Martinez et al. (2015) *Socially-aware robot navigation.*	Physical	Physical	Individual	Robotics
Yeh et al. (2017) *Human-drone interactions.*	Physical	Physical	Individual	Robotics
Pérez et al. (2021) *Development of proxemic mobile apps.*	Physical	Physical	Both	Engineering
Luxey (2019) *Middleware configuration.*	Cyber	Other	Individual	Engineering
Mehta (2020) *Social distancing on human behavior.*	Physical	Physical	Both	Health
Williamson et al. (2021) *Group behavior in a virtual workshop.*	Cyber	Physical	Group	Education
Medeiros et al. (2021) *Safety when using a VR headset.*	Physical	Physical	Individual	Virtual Reality
Gunawan et al. (2021) *Cybercrimes analysis.*	Cyber	Other	Individual	Digital Forensics

Table 4.2: Overview of Existing Research Works Using *Proxemics* for Practical Applications

Reasons for the Selection of *Proxemics* as a Foundation to Calculate Domain-Adaptable Indicators

We hypothesize that by adapting *proxemics* to social media and adapting the five DILMO dimensions, we can establish a foundation for a generic and modular approach to calculating domain-adaptable indicators from social media. This approach could then be used by domain stakeholders to easily build custom indicators to analyze behaviors on social media around their domain. Several factors motivate this choice:

- *Flexibility* (versatility): *proxemics* is versatile and can be adapted to various requirements. Its five dimensions are broad and can be used to model many use cases. From Table 4.2, it can be observed that *proxemics* has been extensively used to model user-oriented similarity measures for a wide range of objectives, including physical metrics like human-drone interactions, behavior analysis in classrooms, picture annotation, social distancing during the COVID-19 pandemic, and other metrics like middleware reconfiguration and cybercrime analysis. These diverse applications highlight the versatility of the *proxemics* theory in addressing various requirements.

- *Domain-Agnostic*: it has no strong correlation to a specific domain; it is a very domain-unaware theory. Table 4.2 shows that existing work leveraging *proxemics* spans vastly different domains of application: Virtual Reality, Health, Video Games, Robotics, Education, Software Engineering, Digital Forensics, and more. This demonstrates the domain-agnostic nature of *proxemics*.

- *Fitness for Social Media*: As shown previously, *proxemics* can be applied to spaces of different natures (e.g., physical space, VR space, social media space) with interacting entities (e.g., real people, video game characters, social media users). As social media platforms can be conceptualized as spaces where various entities interact and maintain distances between each other, it is possible to hypothesize that *proxemics* could be applied to social media and that many aspects of *proxemics* could be naturally linked with social media concepts. For example, *distance* can refer to the distance between social media users, posts, or entities contained in posts such as hashtags, place mentions, etc. Additionally, *location* can refer to the community in which a social media user is positioned in terms of their interests. *Orientation* might signify a sentimental orientation toward certain topics, and *identity* can characterize a social media user by attributes like age, language, etc.

- *Tangible Dimensions*: The proxemic dimensions were originally designed around the physical world. They are practical and tangible, therefore, easier to understand and manipulate. These dimensions can therefore serve as abstractions for more complex concepts. For example, the distance between individuals can be used to represent their level of interaction or social engagement. By leveraging these tangible dimensions, we aim to create similarity measures that simplify the understanding of more abstract or complex phenomena.

We will now explain how we formally redefined the *proxemics* theory for use in social media. This redefinition is a necessary step prior to calculating indicators.

4.3 Formal Redefinition of *Proxemics* in the Context of Social Media

We start by formally adapting the *proxemics* theory through its dimensions (DILMO), see Figure 4.4. This is crucial because *proxemics* is primarily intended to be applied to physical interactions, whether in the physical space (e.g., *a sensor detecting people moving in a room, etc.*) or in cyberspaces (e.g., *the proximity of characters in a video game or a virtual reality world, etc.*). Whereas, in the space of social media, interactions and zones are no longer physical. We will now provide more details on each dimension.

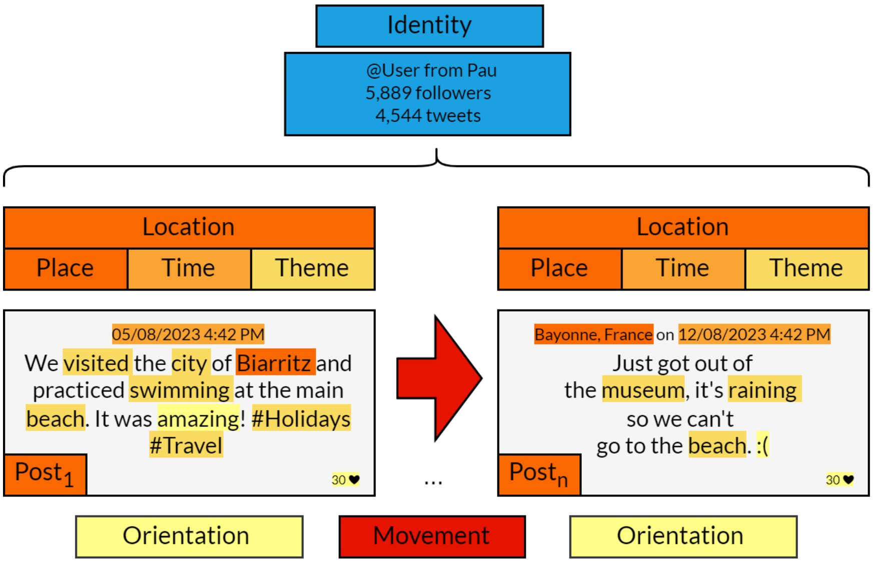

Identity
@User from Pau
5,889 followers
4,544 tweets
Location
Place
Time
Theme
05/08/2023 4:42 PM
We visited the city of Biarritz and practiced swimming at the main beach. It was amazing! #Holidays #Travel
Post1
30 ♥
...
Location
Place
Time
Theme
Bayonne, France on 12/08/2023 4:42 PM
Just got out of the museum, it's raining so we can't go to the beach. :(
Postn
30 ♥
Orientation
Movement
Orientation

Figure 4.4: *Proxemics* Applied to Social Media Posts

4.3.1 Identity Dimension (I)

The *identity* is defined as:

$$identity = (source, user \vee group) \tag{4.1}$$

$source$ is the source (e.g., X/Twitter, Facebook, etc.) from which the $identity$ is sourced. An $identity$ can characterize a single $user$.

$$user = (name, \{\overset{feature_1}{(key_1, value_1)}, \overset{feature_2}{(key_2, value_2)}, \ldots, \overset{feature_n}{(key_n, value_n)}\}) \tag{4.2}$$

$name$ is the username of the associated user. Each user is associated with a set of features, with each feature being modeled as a tuple containing a key and a $value$ (e.g., features can include the number of followers, hometown, nationality, etc.). The number of features depends on what profile information the social media exposes. They can also be computed dynamically based on posts (e.g., category of visitor).

An $identity$ can also characterize a $group$ of users composed of a number n of users. For example, users featuring common features (e.g., users considered influencers because they have reached a certain number of followers).

$$group = (groupName, \{user_1, user_2, \ldots, user_n\}, \{criterion_1, criterion_2, \ldots, criterion_k\}) \tag{4.3}$$

Groups are defined using a set of k criteria. Group criteria are defined according to the specific requirements of the application domain (e.g., what group of users are interesting to analyze for this domain) and the limitations of the social media used (e.g., which variables can be used to define them).

4.3.2 Location Dimension (L)

Posts are defined as:

$$post = (time, \overbrace{\{place_0, place_1, \ldots, place_j\}}^{places}, \overbrace{\{theme_0, theme_1, \ldots, theme_k\}}^{themes}, orientation) \tag{4.4}$$

- $time$ is a timestamp (*e.g., when the post was issued*). For now, we consider one timestamp per post (e.g., the date of the post's publication) for simplicity.
- $\{place_0, place_1, \ldots, place_j\}$ is a set of j places with: $place = (placeUri, placeName)$. Places are instantiated from social media posts' geotags or by extracting named entities (*toponyms*) from their content. We define places by their $placeUri$ in a geographic database (e.g., *OpenStreetMap*) along with their most common toponym $placeName$.
- $\{theme_0, theme_1, \ldots, theme_k\}$ is a set of k themes with:

  $$theme = (conceptUri, conceptName) \tag{4.5}$$

  Themes are domain-specific concepts extracted from the post content according to a semantic resource (e.g., dictionary, thesaurus, or ontology) defining the domain of interest. Each theme has a unique $conceptUri$ referencing the semantic resource and a $conceptName$.

4.3.3 Orientation Dimension (O)

The *orientation* is defined by $orientation = (sentiment, engagement)$.
$sentiment$ represents the polarity of the dominant sentiment expressed in the associated post:

$$sentiment \in \{positive, negative, neutral\} \tag{4.6}$$

$engagement$ is the engagement orientation with:

$$engagement = (engagement\_metric_1, engagement\_metric_2, \ldots, engagement\_metric_n) \tag{4.7}$$

The tuple $(engagement\_metric_1, engagement\_metric_2, \ldots, engagement\_metric_n)$ can vary depending on the social media platform used. For example, in *X/Twitter*, we consider the number of likes, retweets, quotes, and answers, so: $engagement = (likes, retweets, quotes, answers))$

4.3.4 Movement Dimension (M)

We define a social media *movement* (*multidimensional trajectory*) of size n as a tuple with:

$$movement = (identity, \langle post_0, post_1, \ldots, post_n \rangle) \tag{4.8}$$

A $movement$ is a sequence of posts $\{post_0, post_1, \ldots, post_n\}$ associated with an $identity$. Each $movement$ can be broken down into sub-movements (dimensional trajectories).

The *spatial trajectory* ($movement_{spatial}$) is the sequence of places identified in a $movement$. It provides a comprehensive overview of the places an $identity$ has mentioned. It is a sequence of tuples because each post can be associated with any number of places:

$$movement_{spatial} = \tag{4.9}$$
$$\langle \overset{post_0}{(place_0, place_1, \ldots, place_n)},\ \overset{post_1}{(place_0, place_1, \ldots, place_n)},\ \ldots,\ \overset{post_n}{(place_0, place_1, \ldots, place_n)} \rangle$$

The *temporal trajectory* ($movement_{temporal}$) is the sequence of timestamps identified in a $movement$:

$$movement_{temporal} = \langle \overset{post_0}{time_0},\ \overset{post_1}{time_1},\ \ldots,\ \overset{post_n}{time_n} \rangle \tag{4.10}$$

The *thematic trajectory* ($movement_{thematic}$) is the sequence of themes identified in a $movement$. It highlights the domain-specific concepts associated with the $identity$ (e.g., the activities practiced by a certain category of visitors). It is a sequence of tuples (each post can be associated with any number of themes):

$$movement_{thematic} = \tag{4.11}$$
$$\langle \overset{post_0}{(theme_0, theme_1, \ldots, theme_n)},\ \overset{post_1}{(theme_0, theme_1, \ldots, theme_n)},\ \ldots,\ \overset{post_n}{(theme_0, theme_1, \ldots, theme_n)} \rangle$$

The *sentimental trajectory* ($movement_{sentimental}$) provides the sequence of sentiments for a given $identity$, it can be considered as a sentimental trajectory.

$$movement_{sentimental} = \langle \overset{post_0}{sentiment_0},\ \overset{post_1}{sentiment_1},\ \ldots,\ \overset{post_n}{sentiment_n} \rangle \tag{4.12}$$

These movements (minus the temporal one) are **ubiquitous**, because in any given $post$, an $identity$ can be in several locations at the same time.

4.3.5 Distance Dimension (D)

Static distances $distance(e_1, e_2)$ can be calculated between social media entities e of the same type. Indeed, we rely on existing metrics for the *Distance* dimension, that work only between entities of the same nature. We will explain later (see Section 4.6) how these distances are calculated. We consider the following types of entities e (*users, groups, places, periods, or themes*).

$$distance(e_1, e_2) \text{ with } e_1, e_2 \in \{identity, place, time, theme\} \text{ and } e_1 \text{ is of the same type as } e_2 \tag{4.13}$$

This formal redefinition allows us to design the APs Trajectory Model (Masson et al., 2023b), which leverages this formal redefinition of *proxemics* to model movements and interactions on social media in a generic and domain-independent manner. It will be used as support later in the thesis to calculate domain-adaptable indicators.

4.4 The APs Trajectory Data Model

Here, we introduce the APs Trajectory Model and its associated OCL constraints[1]. It is designed in five parts, following the core dimensions of *proxemics* (DILMO, see Table 4.1).

[1] Object Constraint Language

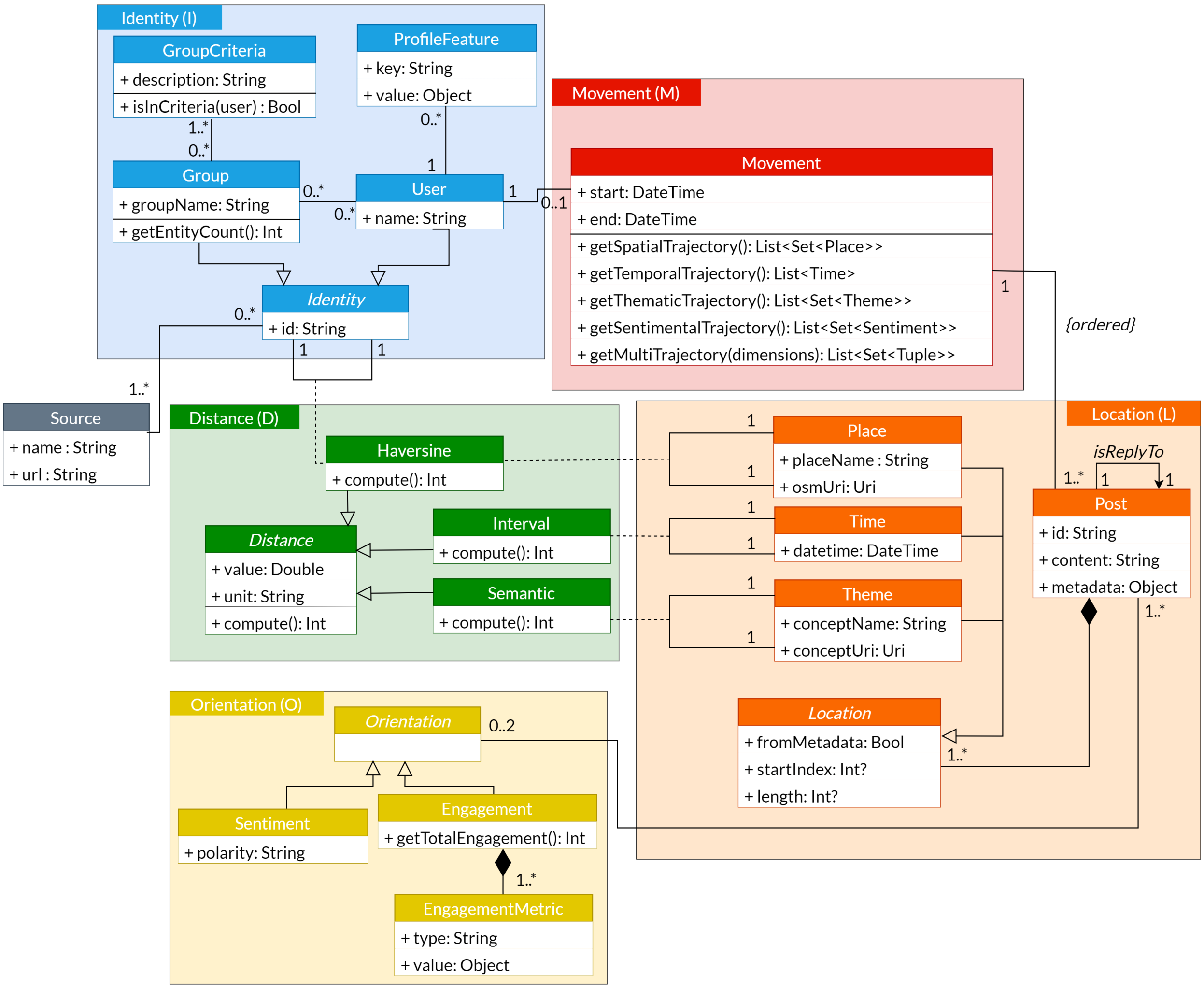


Figure 4.5: UML Class Diagram of the APs Trajectory Model — Designed Around the Five Dimensions of *Proxemics* (DILMO)

The APs Framework (see Figure 1.5) is designed to be compatible with any type of social media data, as long as it conforms to this model, thus ensuring its applicability across various domains and social media platforms. Figure 4.5 presents the UML[2] class diagram of the model, which is instantiated step-by-step throughout the framework's pipeline.

This model is (1) *multidimensional* (designed around the five dimensions of *proxemics*), (2) *modular* (the use of all five dimensions is not mandatory; one can combine any number of them), and (3) *extensible* (many classes are abstract and methods are virtual, allowing users to provide new implementations) and (4) *generic* (this genericity extends to both the social media source and the domain of application).

4.4.1 Identity Dimension (I)

The Identity (I) dimension allows for the modeling of the studied population: individual users (along with their profile features) or user groups. A user always has a single movement associated with them (*based on the posts issued by them*) whereas, in the case of groups, it depends on the number of users making up the group. Groups are defined according to criteria applied to profile

[2] Unified Modeling Language

features (e.g., an influencer group could be defined according to the amount of followers associated with users).

4.4.2 Movement Dimension (M)

The Movement (M) dimension provides the ordered sequence of posts belonging to a given user. It gives a comprehensive view of a user's activities on the chosen social media and allows linking posts together to create multidimensional trajectories. Additionally, it can be broken down into several sub-trajectories (*spatial, thematic, spatio-thematic, tempo-sentimental, etc.*), see below:

```
context Movement::getSpatialTrajectory() :
self.posts->collect(postItem | postItem.locations->
select(o | o.oclIsTypeOf(Place))->asSet())->asSequence()
```

```
context Movement::getTemporalTrajectory() :
self.posts->collect(postItem | postItem.locations->
select(o | o.oclIsTypeOf(Time)))->asSequence()
```

```
context Movement::getThematicTrajectory() :
self.posts->collect(postItem | postItem.locations->
select(o | o.oclIsTypeOf(Theme))->asSet())->asSequence()
```

```
context Movement::getSentimentalTrajectory() :
self.posts->collect(postItem | postItem.orientation->
select(o | o.oclIsTypeOf(Sentiment))->asSet())->asSequence()
```

4.4.3 Location Dimension (L)

The Location (L) dimension models the posts themselves along with their associated *locations*. We move away from the solely physical definition of *location* and consider three types of locations. These can be (1) spatial (places based on toponyms extracted from posts or geotag metadata), (2) temporal (timestamps), or (3) thematic (domain-specific concepts aligned with a semantic resource). Themes are defined according to the studied domain's description (domain-specific ontology, thesaurus, or dictionary). These semantic resources provide additional hierarchy information. When it comes to places, they are associated with a unique identifier linked to a spatial database. This allows for featuring relationships (e.g., a municipality is within a region, itself within a country). A given *post* can be in several *locations* at the same time, making the model *ubiquitous* (see Figure 4.4). From now on, and until the end of this chapter, when we reference the term *location*, it will refer to this proxemic locations (which, as mentioned before, can be themes, places, or periods), and not a necessarily spatial locations.

4.4.4 Orientation Dimension (O)

The Orientation (O) dimension contains contextual and enrichment data, such as the sentiment of the associated post (positive, negative or neutral) and the engagement associated with it (e.g., number of replies, likes and quotes). The classes for both locations and orientations are designed to be extensible, thus providing the flexibility to incorporate new classes as desired (e.g., when studying politics, one could imagine a *political orientation* dimension). Unlike locations, a given post can only have one orientation of each type (see OCL constraint below).

```
context Post
inv : self.orientation ->
select (o|o.oclIsTypeOf(Sentiment)) -> size() <= 1
and
inv : self.orientation ->
select (o|o.oclIsTypeOf(Engagement)) -> size() <= 1
```

4.4.5 Distance Dimension (D)

Lastly, the Distance (D) dimension helps in modeling and storing static distances (physical distance, thematic distance in the semantic resource, temporal intervals, profile-based distance) between entities of the same type, specifically between (1) two identities (*user* or *group*) as well as between (2) two locations (*place*, *time* or *theme*). We will not go into details on this dimension for now, as it will be detailed later, in Section 4.6.

4.4.6 Model Instantiation in the APs Framework

We have decided to present the APs Trajectory Model in this chapter due to its link to *proxemics*, but the model is used during all steps of the framework as depicted in Figure 4.6. This figure shows the same model's instantiation process throughout the APs Framework.

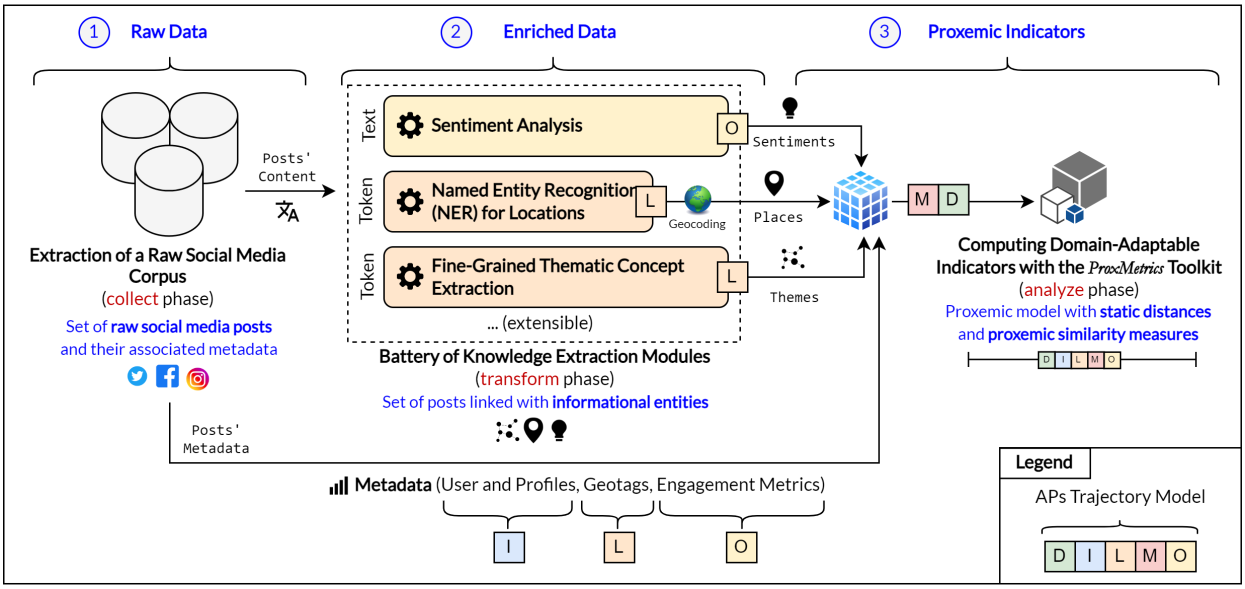

1 Raw Data
2 Enriched Data
3 Proxemic Indicators
Posts' Content
Extraction of a Raw Social Media Corpus
(collect phase)
Set of raw social media posts and their associated metadata
Posts' Metadata
Text
Sentiment Analysis
O
Sentiments
Token
Named Entity Recognition (NER) for Locations
L
Geocoding
Places
Token
Fine-Grained Thematic Concept Extraction
L
Themes
... (extensible)
Battery of Knowledge Extraction Modules
(transform phase)
Set of posts linked with informational entities
M D
Computing Domain-Adaptable Indicators with the ProxMetrics Toolkit
(analyze phase)
Proxemic model with static distances and proxemic similarity measures
D I L M O
Metadata (User and Profiles, Geotags, Engagement Metrics)
I
L
O
Legend
APs Trajectory Model
D I L M O

Figure 4.6: Overview of the Instantiation Process of the APs Trajectory Model

The *Collect* phase instantiates the model (Figure 4.6, *Raw Data*), which is a raw, unstructured social media corpus. At this point, the model contains only the raw posts and the users' profiles. In the *Transform* phase (Figure 4.6, *Enriched Data*), knowledge extraction modules are applied to the posts' content to categorize them by sentiment (*Orientation*) and extract informational entities such as places (we geocode them too) and thematic concepts, linked with the semantic resource used (Figure 4.1, *Semantic Domain Description*) (*Location*). In Chapter 3, we determined the best techniques for these three knowledge extraction tasks in the tourism domain. This *Transform* phase allows us to instantiate the multidimensional trajectories (*Movement*). Finally, in the *Analyze* phase

(Figure 4.6, *Proxemic Indicators*), static distances are computed on the model (*Distance*) along with domain-adaptable indicators expressed as proxemic similarity measures. We will discuss this later, in Section 4.6.

Having established a foundation for the modeling and storage of social media entities and interactions that are not bound to specific domains and sources, our next step involves the proposition of a toolkit to calculate domain-adaptable indicators from this generic model. As mentioned in Section 4.1, we have decided to express these indicators as similarity measures, specifically proxemic similarity measures. These are indicators calculated between two social media entities, such as users, and groups, or informational entities such as places, themes, and periods. We will now review existing similarity measures that we could use and extend to design proxemic-based ones.

4.5 Related Work: Similarity Measures and Criteria Combination

Before going further, we review existing similarity measures (see Subsection 4.5.1) and the main techniques for combining several criteria (see Subsection 4.5.2). Indeed, as we aim to be domain-adaptable, our proxemic-based similarity measures should be composite and customizable by users.

4.5.1 Existing Similarity Measures

Social media platforms have become a focal point for the research and application of various algorithms (Anderson et al., 2012), especially with regard to trend analysis (Bhor et al., 2018), content recommendation (Jiang and Yang, 2017), event detection (Becker et al., 2010; Huang et al., 2021), and ad targeting (Knoll, 2016). One of the key components for these tasks is the ability to assess similarity between entities, whether these items are words, posts, users, or media. We have chosen to divide existing similarity measures into four core families (traditional, series-based, graph-based, and deep learning).

Table 4.3 provides a side-by-side comparison of existing similarity measures aligned with the various types of social media data contained in the APs Trajectory Model and the headers' colors correspond to the proxemic dimension of the data in the model (refer to Figure 4.5 for details). We consider that a user's features can be either categorical (*e.g.,* gender, occupation, etc.) or numerical (*e.g.,* age, height, etc.). We have specified with the *sets* keyword when the similarity is calculated between unordered sets of entities (rather than individual ones), while the *seq* keyword indicates ordered sequences. The ✓ symbol indicates that the measure is compatible and fits the given data type, requiring little to no modification. The ✳ symbol indicates that the measure could be adapted but would require substantial redesign or modification. The absence of a symbol suggests that the measure is not fit or that adapting the measure would require extensive work.

The underlined measures are those we selected to adapt. We hypothesized that we could rely on these measures as a foundation to build a proxemic similarity toolkit and adapt them to assess similarity based on proxemic dimensions. These similarity measures are widely used, cover most of the data types we are dealing with, are less complex than graph or series-based ones, and do not necessitate time-consuming or costly training datasets.

	Profiles' Features		Posts' Features				
	Numerical	Categorical	Place	Time	Theme	Sentiment	Engagement
Example	*25 y.o.*	*Male*	*43.47, -1.41*	*2023-10-09*	*Natural::Beach*	*Positive*	*128 likes*
Traditional							
Euclidean (Johansson et al., 2013)	✓		✳	✳	✳	✓	✓
Manhattan (Wang et al., 2016b)	✓		✳	✳	✳	✓	✓
Minkowski (Groenen et al., 1995)	✓		✳	✳	✳	✓	✓
Haversine (Nguyen et al., 2017)			✓				
Mahalanobis (Leys et al., 2018)	✓			✳	✳	✳	✳
Jaccard (Zangerle et al., 2013)		✓ (set)		✳ (set)	✓ (set)		
Pearson (Sponcil and Gitimu, 2013)	✓			✓ (seq)			
Dice (Duarte et al., 1999)		✳ (set)		✳ (set)	✓ (set)	✳	
Cosine (Lahitani et al., 2016)	✓			✳ (seq)			
Hamming (Bookstein et al., 2002)		✳		✳ (seq)		✓ (seq)	
Levenshtein (Navarro, 2001)		✓		✓ (seq)	✓ (seq)		
Chebyshev (Coghetto, 2016)	✓	✳		✳ (seq)	✳ (seq)		
Earth Mover (Rubner et al., 2000)	✓		✳ (seq)	✳ (seq)			✳
Semantic (Wu and Palmer, 1994)					✓		
Series-based							
DTW (Müller, 2007)			✳ (seq)	✓ (seq)			
TraFoS (Varlamis et al., 2021)			✓ (seq)				
Hausdorff (Huttenlocher et al., 1993)							
Frechet (Alt and Godau, 1995)			✓ (seq)				
TRACLUS (Jiashun, 2012)			✓ (seq)				
LCSS (Bergroth et al., 2000)			✓ (seq)	✓ (seq)	✓ (seq)		
CED (Moreau et al., 2020)					✓ (seq)		
Graph-based							
Node Similarity (Tang et al., 2016)	✓	✓	✳	✳	✳		
Random Walk (Xia et al., 2019)	✳	✳	✳	✳	✳		
Deep Learning							
Word Embeddings (Liu et al., 2015)					✓	✳	
User Embeddings (Amir et al., 2016)	✓	✓					
Pretrained Models (Devlin et al., 2019)	✳	✳	✳	✳	✳	✳	✳

Table 4.3: Comparison of the Applicability of Existing Similarity Measures for Social Media Data

These include the *Haversine* (Nguyen et al., 2017; Baucom et al., 2013) and *Euclidean* (Johansson et al., 2013) distances to measure the straight-line distance between two coordinates, the first taking into account the curvature of the Earth, which is essential for dealing with spatial coordinates such as geocoded toponyms or posts' geotags. The *Jaccard Index* assesses the similarity of two sets of data and is applicable for evaluating social media trends like hashtag usage (Zangerle et al., 2013). Lastly, the semantic (*Wu-Palmer*) similarity determines the closeness of semantic concepts in a taxonomy, it calculates similarity by considering the elements' depths in the taxonomy and the depth of their *Least Common Subsumer* (LCS), making it particularly useful for semantic similarity measures (Wu and Palmer, 1994).

We will now examine how similarity metrics can be combined to create composite ones, which is crucial in our use case, given that we have five distinct proxemic dimensions (DILMO).

4.5.2 The Challenge of Criteria Combination

The diversity of social media data often requires a nuanced approach to measuring similarity. A single similarity measure may not adequately address the multifaceted nature of the data. Therefore, the combination of various similarity measures can provide a more robust and comprehensive understanding. This section outlines strategies for combining multiple measures to improve indicators' accuracy in various domains.

- In *healthcare*, criteria such as patient history, symptom severity, and laboratory test results are often combined. While weighted means are common, other methods like decision trees or Bayesian networks might be used, depending on the complexity and nature of the data. For example, Khan et al. (2008) used fuzzy decision trees to combine various biological indicators for disease prediction.

- *Financial institutions* might use weighted means, logistic regression, or machine-learning models to combine criteria like credit history, current debts, and income levels for risk assessment. The approach chosen often depends on the requirement for interpretability compared to predictive accuracy. For example, Bolton et al. (2010) explored the application of logistic regression in credit scoring models.

- In *education*, combining criteria such as student engagement, performance metrics, and feedback can involve weighted means, cluster analysis, or even neural networks. The choice depends on the educational context and the specific goals of the analysis. For instance, Ng et al. (2016) employed cluster analysis to categorize student motivation and learning behaviors based on multiple metrics.

- Lastly, *environmental scientists* often rely on spatial analysis techniques to combine criteria like pollutant levels, biodiversity indices, and land use patterns. The method chosen often reflects the scale and complexity of the environmental data. A study by Lu et al. (2015) leveraged spatial analysis to assess the impacts of various environmental factors on ecosystem health.

In our proxemic similarity toolkit, we plan to first experiment with the weighted mean. This decision is driven by three main factors. Firstly, as demonstrated above, various domains employ distinctly different methods for combining indicators. Therefore, to ensure domain independence, it is crucial for us to adopt a combination technique that can be flexibly adapted to each domain's specific requirements and that is not too specialized. Secondly, we aim to integrate the expertise of domain stakeholders into the measure. Stakeholders should have the ability to adjust the impact of each dimension to tailor the indicators to their unique requirements. Thirdly, we require a combination technique that is easily interpretable by users who are not computer scientists. A weighted mean, in contrast to more complex combination methods such as machine-learning models, provides a clear and understandable rationale for the assigned weights of different measures. We will now explain how the APs Trajectory Model (see Figure 4.5) and the similarity measures we have selected are used in our proxemic similarity toolkit.

4.6 *ProxMetrics*: Modular Toolkit to Evaluate Proxemic Similarity in Social Media

We introduce *ProxMetrics* (Masson et al., 2024b), a modular toolkit designed to assess the similarity between multidimensional entities on social media based on the five dimensions of *proxemics* (DILMO). This toolkit is designed to be modular and generic, adaptable to any social media platform, and flexible to accommodate a wide variety of user requirements.

4.6.1 Proxemic Entity Definitions

We define *proxemic similarity* on social media as the perceived relational closeness or association between entities on social media, based on the nature, frequency, and depth of their interactions or mentions within social media platforms. Various factors can be considered when evaluating proxemic similarity, including analysis of individual posts, user trajectories, engagement levels, sentiment analysis, or even profile information. In this work, we consider the following categories of social media entities (which will serve as proxemic references). Firstly, we have *dynamic entities* (derived from the *Identity* dimension in Figure 4.5): *users* and *groups*. They actively *interact* and *move* within the landscape of social media, analogous to people in physical *proxemics*.

- *Users* ($user \equiv$ 👤): individual social media users, whether physical (e.g., *a person*) or corporate, institutional ones (e.g., *an institution*, *a company*).
- *Groups* or *demographics* ($group \equiv$ 👥 $= \{user_1, user_2, \ldots, user_n\}$): groups of social media users. They are defined according to the domain of study and can be based on shared features (e.g., *French users*, *influencers*, *foreign visitors*, *etc.*).

Secondly, there are *static entities* (derived from the *Location* dimension in Figure 4.5): *places*, *periods*, and *themes*. These are extracted from users' posts and, unlike dynamic entities, do not interact on their own. Instead, they *appear* in the user posts. These static entities are analogous to objects in physical *proxemics*.

- *Places* or *spatial entities* ($place \equiv$ 📍): places mentioned on social media. Different levels of granularity are possible, such as points of interest, districts, municipalities, regions, countries, etc. Place can be extracted from the posts' metadata (*geotags*) or from the content of the posts.

- *Periods* or *temporal entities* ($time \equiv$ 🕒): periods associated with social media posts. Different levels of granularity are possible, including hour, day, week, month, year, season, day of the week, etc. They are extracted from the posts' timestamps (metadata).
- *Themes* or *thematic entities* ($theme \equiv$ 📖): domain-specific thematic concepts mentioned on social media. They are assigned to a semantic resource (e.g., *dictionary, thesaurus, ontology*). Different levels of granularity, like levels within a thesaurus or ontology, are possible. They are extracted from the posts' content.

4.6.2 Proxemic Similarity Definition

In *proxemics*, the selection of a *reference entity*, or the *center entity* (E_{ref}), is essential (refer to Subsubsection 4.2.2). While the traditional physical context of *proxemics* mostly uses individuals or sensors as references (whose interactions are under observation), the landscape of social media provides a more diverse range of entities as potential references. This could be, for example, a specific user ($user$), a group ($group$), a place ($place$), a theme ($theme$), or a timestamp ($time$).

$$E_{ref} \in \{user, group, place, theme, time\} \tag{4.14}$$

Proxemic similarity (P_s) is measured between a chosen reference entity E_{ref} and a set of target entities denoted as τ. This relationship gives rise to a wide array of potential entity pairings (25 combinations), including *user to users, user to groups, place to users, place to places*, among others.

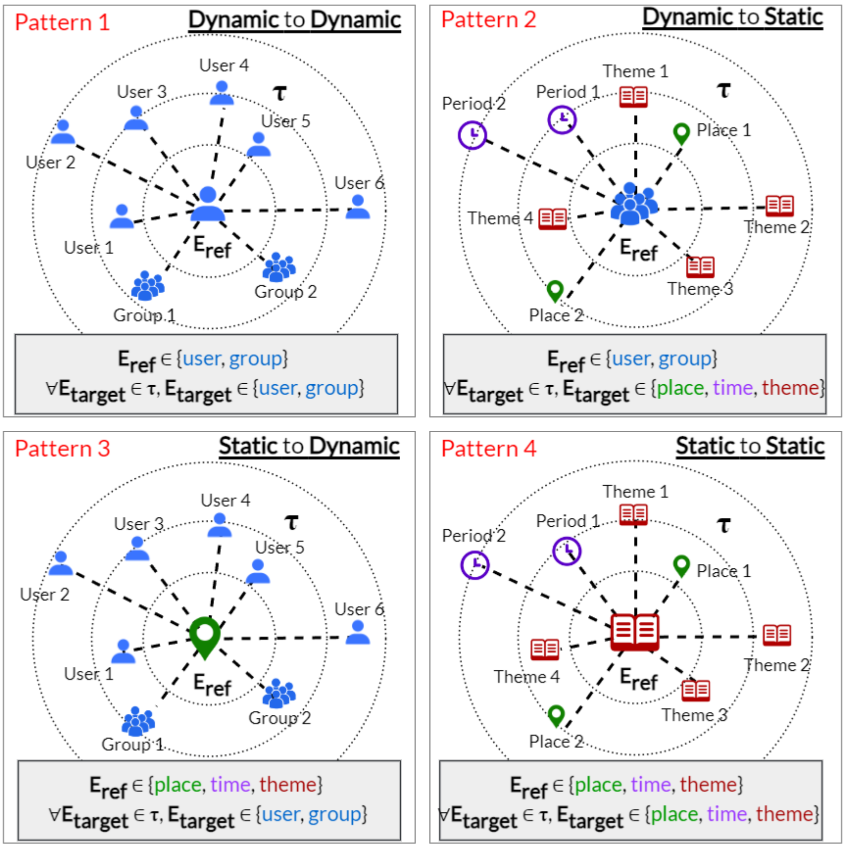


Figure 4.7: The Four Proxemic Similarity Patterns in the *ProxMetrics* Toolkit

To systematically categorize these pairings, we have identified four primary proxemic similarity patterns, as illustrated in Figure 4.7. The selection of the four patterns is driven by the fundamental differences between static, informational entities present in posts, and dynamic entities common on social media platforms, necessitating distinct approaches for evaluating their similarity.

In the center of the proxemic reticle is the reference entity E_{ref}. Surrounding it is a set of target entities τ. We denote entities in this set as E_{target} (with $E_{target} \in \tau$). The visual distance between these and the reference entity represents their relative proxemic similarity. This distinction in pattern is necessary because, unlike in physical *proxemics*, determining the proxemic similarity between dynamic entities (users or groups issuing posts), static entities (informational entities found in posts), or a combination of both cannot be done in the same manner.

Depending on their level of proxemic similarity in regard to the reference entity E_{ref}, target entities τ might be categorized into different *proxemic zones*. These zones are demarcated by the light gray dashed lines in Figure 4.7. The number and range of proxemic zones are determined based on the study domain and by the end-users. For example, in the tourism domain, they might represent the degree of attraction a visitor feels towards nearby POIs (Points of Interest). Unlike in traditional physical *proxemics* (with intimate, personal, social, and public zones), there are no universal definitions of zones when it comes to social media. We will now explain how we assess the similarity between entities.

4.6.3 Proxemic Similarity Design

The formula for proxemic similarity (P_s) is grounded in the five proxemic dimensions (Distance, Identity, Location, Movement, Orientation), as detailed in Figure 4.8.

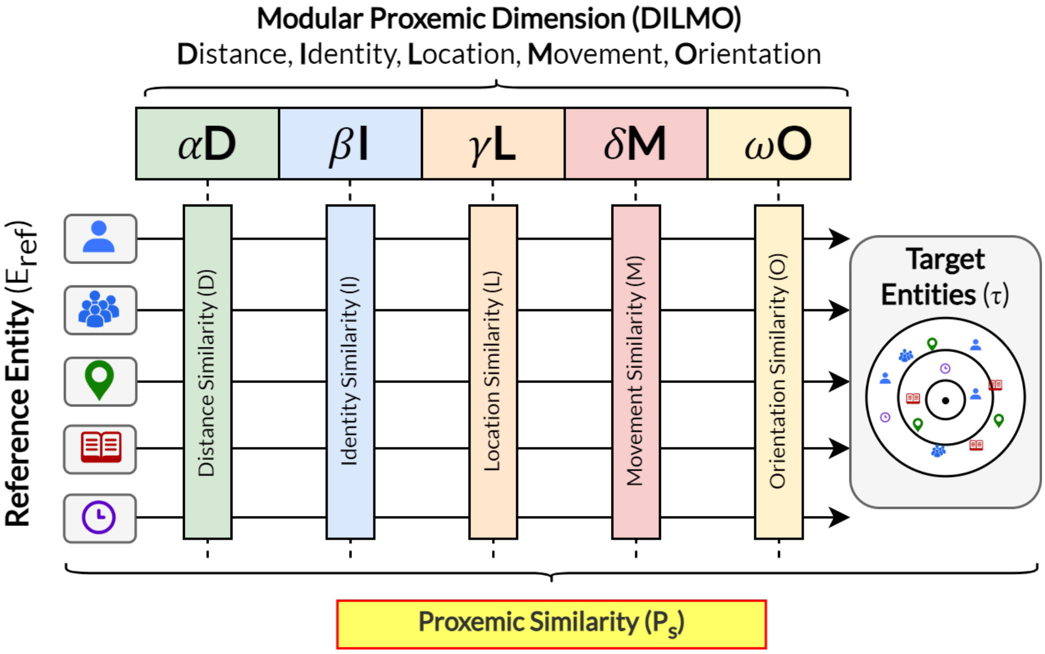


Figure 4.8: Overview of the *ProxMetrics* Toolkit

It is therefore composite and modular; the user can modulate (amplify or reduce) the impact of certain dimensions depending on the requirements he wishes to address. This is done through five coefficients (here α, β, γ, δ, ω). As a reminder, we require a composite and modular similarity based on proxemic dimensions because we need to be domain-independent and adaptable to various requirements. The same toolkit should be usable across a wide variety of different domains with multiple requirements and the results must be interpretable by users who are not necessarily computer scientists.

Given a reference entity E_{ref} and a target entity E_{target}, we introduce a composite measure $P_s(E_{\text{ref}}, E_{\text{target}})$. This measure is an aggregation of five sub-measures, each corresponding to one of the proxemic dimensions (DILMO). These sub-measures quantify the similarity between the 2 entities across each dimension, providing a quantitative assessment of their proxemic relationship.

Proxemic similarity (P_s) is expressed as a numerical value ranging from 0 to 1. A value of 1 indicates a strong proxemic similarity, while a value of 0 points to a lack of similarity.

$$\begin{gathered}
P_s(E_{ref}, E_{target}) = \alpha D(E_{ref}, E_{target}) + \ \beta I(E_{ref}, E_{target}) + \ \gamma L(E_{ref}, E_{target}) \\
+ \ \delta M(E_{ref}, E_{target}) + \ \omega O(E_{ref}, E_{target}) \\
\text{with } \alpha + \beta + \gamma + \delta + \ \omega = 1 \text{ and} \qquad (4.15) \\
0 \leq P_s, D(E_{ref}, E_{target}), I(E_{ref}, E_{target}), L(E_{ref}, E_{target}), M(E_{ref}, E_{target}), O(E_{ref}, E_{target}) \leq 1
\end{gathered}$$

We hypothesize that by blending and modulating these five unidimensional measures, we can create domain-specific indicators. This approach aims to address domain requirements in a manner that is *generic*, functioning across various social media platforms, *domain-independent*, applicable to different domains of application, and *versatile*, capable of accommodating a wide range of end-user requirements.

It is important to note that this toolkit is not intended for direct use by non-computer scientists. Therefore, it is not to be manipulated directly by novice users (in our case, domain stakeholders). Instead, it must be preliminary parameterized by trained users who create various configurations according to domain requirements. The results are then presented to novice end users. We will elaborate on that in Chapter 5. For now, let's present how each component is calculated.

4.6.4 Distance Similarity: $D(E_{ref}, E_{target})$

As previously mentioned, we opted to align with physical *proxemics* for the Distance $D(E_{ref}, E_{target})$. For simplicity, we currently allow the *Distance* dimension to be applied only when calculating the proxemic similarity between two entities of the same kind (e.g., two users, two places, two themes, two periods, etc.). This choice was made so we can leverage already established and efficient similarity measures for this dimension, which can be calculated only between entities of the same nature (e.g., two places, two themes, two dates, etc.).

For users and places, we leverage the *Haversine* distance (Nguyen et al., 2017; Baucom et al., 2013) (denoted as $D_{physical}$), which measures the straight-line distance between two points while accounting for the Earth's curvature. This decision is grounded in the utility of physical metrics in social media contexts. For example, it helps determine whether two social media users are in close proximity or if two municipalities are in the same region. This physical metric could be useful in recommendation scenarios, where we want to recommend POIs (*Points of Interest*) to users that are physically close to them. In cases where the real time physical positioning of a user is not supported by the social media platform, we rely on their most recently recorded physical geolocation.

For user groups, we calculate the centroid position of all members, indicating the group's predominant geolocation.

For themes, we evaluate their semantic similarity (denoted as $D_{semantic}$) using the *Wu-Palmer* methodology (Wu and Palmer, 1994), which determines similarity based on their least common subsumer (LCS). This method helps ascertain whether two themes are semantically *close* (e.g., *Beach* with *Sea*) or *far* (e.g., *Beach* with *Museum*).

For dates, we evaluate the interval ($D_{interval}$) in hours between them, and for periods, we reference the median date. This enables us to determine the temporal proximity of dates or periods, determining whether they occurred close or not.

Lastly, it is important to note that the *Haversine* distances and time intervals undergo normalization between 0 and 1 using min-max normalization. Normalization parameters require to be tweaked depending on the spatial area and time range covered by the social media corpus in use.

4.6.5 Identity Similarity: $I(E_{ref}, E_{target})$

The *Identity* dimension $I(E_{ref}, E_{target})$ uses profile features to calculate the similarity. Social media users possess various profile features such as age, gender, occupation, and more. These features are crucial for understanding behavior by emphasizing the unique characteristics of individual profiles. The primary goal of this dimension is to bridge demographic differences and to detect similar groups or users. If we refer back to the patterns illustrated in Figure 4.7, here are general use cases of how this dimension is applied along with examples in the tourism domain:

- *Pattern 1* $E_{ref} = user \vee group, E_{target} = user \vee group$ Detection of users or groups with similar profiles based on their features (e.g., user or group similarity based on their age, language, gender, etc.). For example, this could be used in a visitors' connection system.

- *Pattern 2* $E_{ref} = user \vee group, E_{target} = place \vee time \vee theme$ Recommendation of places, themes or periods to users or groups based on places visited, themes mentioned or posting periods by users with a similar profile. For example, older users may prefer certain touristic activities more (demographic filtering-based recommendations).

- *Pattern 3* $E_{ref} = place \vee time \vee theme, E_{target} = user \vee group$ Detection of which users or groups primarily visit a given place, are interested in a certain theme or are active during a certain period of the day or year. For example, this could be used to compute indicators for tourism professionals, such as understanding the typical profile of visitors visiting a given Point of Interest (POI).

- *Pattern 4* $E_{ref} = place \vee time \vee theme, E_{target} = place \vee time \vee theme$ Comparison of themes, places, or periods based on the profile of users or groups mentioning them. For example, two places (e.g., shop, restaurant, etc.) might want to identify how similar their visitors are.

Pattern 1

Let's begin by defining I between two individual users (*Pattern 1*). Given two users E_{ref} and E_{target}, the component $I_{individual}(E_{ref}, E_{target})$ is defined as:

$$I_{individual}(E_{ref}, E_{target}) = \frac{\sum_{i=1}^{n} w_i \times s_{feature}(E_{ref}^i, E_{target}^i)}{\sum_{i=1}^{n} w_i}$$
$$\text{with } E_{ref} = user \text{ and } E_{target} = user \tag{4.16}$$

- n represents the number of profile features considered (e.g., age, number of followers, nationality, etc.).
- w_i represents the weight of the i^{th} feature.
- E_{ref}^i is the i-th feature of E_{ref} (here E_{ref} refers to a user).
- $s_{feature}(E_{ref}^i, E_{target}^i)$ denotes the similarity between the i^{th} feature of users E_{ref} and E_{target} respectively.

Let's now explain how we calculate $s_{feature}$. For numerical features (*e.g., age, height*), we use the normalized Manhattan distance (Wang et al., 2016b):

$$s_{feature}(E_{ref}^i, E_{target}^i) = \frac{|E_{ref}^i - E_{target}^i|}{\text{max}_{\text{attr}} - \text{min}_{\text{attr}}} \tag{4.17}$$

- max_{attr} and min_{attr} are the maximum and minimum possible values for the feature respectively (for example, for age, we would consider the maximum range of ages featured in our dataset of users).

For categorical features (*e.g., gender*):

$$s_{feature}(E_{ref}^i, E_{target}^i) = \begin{cases} 0 & \text{if } E_{ref}^i = E_{target}^i \\ 1 & \text{if } E_{ref}^i \neq E_{target}^i \end{cases} \tag{4.18}$$

We can now extend this to groups of users, given two groups of users E_{ref} and E_{target} (also *Pattern 1*), the average similarity $I_{group}(E_{ref}, E_{target})$ is:

$$I_{group}(E_{ref}, E_{target}) = \frac{1}{m \times p} \sum_{i=1}^{m} \sum_{j=1}^{p} I_{individual}(E_{ref}^i, E_{target}^j)$$
$$\text{with } E_{ref} = group \text{ and } E_{target} = group \tag{4.19}$$

- m is the number of profiles in group E_{ref}.
- p is the number of profiles in group E_{target}.
- E_{ref}^i is the i^{th} user in group E_{ref}.
- E_{ref}^j is the j^{th} user in group E_{target}.

- $I_{individual}(E^i_{ref}, E^j_{target})$ is the similarity between the i^{th} user in E_{ref} and the j^{th} user in E_{target} computed using the individual formula described before.

When calculating the similarity between a user and a group, we assimilate the user to a group of size 1 (e.g., containing only the user) and therefore use the I_{group} formula.

Pattern 2, 3 and 4

When one of the entities considered (E_{ref} or E_{target}) is a static entity (e.g., themes, places, or periods), which is the case in *Pattern 2*, *Pattern 3*, and *Pattern 4*, we assimilate it to the subset of users who have referenced it in their posts.

For example, assessing the *identity similarity* between theme$_1$ and theme$_2$ involves computing a group similarity I_{group} between the set of users who have mentioned theme$_1$ in their posts and the set of users who have mentioned theme$_2$ in their posts. This enables us to assess whether these entities are associated with similar users' demographics. In the case of *Pattern 4*, this would translate as I_{group} being called like this $I_{\text{group}}(\text{getUsersMentioning}(E_{\text{ref}}), \text{getUsersMentioning}(E_{\text{target}}))$ where getUsersMentioning(static_entity) returns the list of users having mentioned the static entity.

4.6.6 Location Similarity: $L(E_{ref}, E_{target})$

The *Location* dimension $L(E_{ref}, E_{target})$ operates differently based on the pattern of proxemic similarity in use (see Figure 4.7). Specifically, it has three variations.

Pattern 1

For *Pattern 1*, which covers four proxemic similarity entity pairing (user to users, user to groups, group to users, and group to groups), we adapt the *Jaccard* Index (Zangerle et al., 2013) used to compute the similarity between sets of places, periods, and themes. We evaluate the co-occurrences of locations between E_{ref} and E_{target}. The more locations (spatial, temporal, and thematic) two users or groups share, the more similar they are considered to be. For example, if two users frequently mention visiting the same municipalities or attending the same events, they are considered to have a high location similarity. Additionally, we want locations mentioned in recent posts to weigh more (we consider that, as they are more recent, they are more relevant to domain stakeholders). Therefore, we introduce a time decay factor $w_{freshness}$, which is based on the freshness of posts (e.g., a post issued today will weight 1, while a post issued x days ago will weigh less).

$$w_{freshness}(post) = e^{-\lambda \times (currentDate - post.date)} \tag{4.20}$$

- $(currentDate - post.date)$ is the difference in days between the current date and the post's issuance date.

- λ is a constant that controls the decay rate (how fast older locations are weighted less). It must be tweaked based on the time range covered by the data. We will use $\lambda = 0.01$ for a balanced effect in our tourism dataset (spanning three months).

The time-weighted locations' similarity score between users or groups could be defined as:

$$\begin{gathered}L_{individual}(E_{ref}, E_{target}) = \\ \sigma \times L_{spatial}(E_{ref}, E_{target}) + \theta \times L_{temporal}(E_{ref}, E_{target}) + \zeta \times L_{thematic}(E_{ref}, E_{target}) \\ \text{with } \sigma + \theta + \zeta = 1 \text{ and } E_{ref} \in \{user, group\} \text{ and } E_{target} \in \{user, group\}\end{gathered} \tag{4.21}$$

- σ, θ, ζ are coefficients used to modulate the strength of each type of location (*spatial, temporal, thematic*).
- $L_{spatial}(E_{ref}, E_{target}), L_{temporal}(E_{ref}, E_{target}), L_{thematic}(E_{ref}, E_{target})$ are individual locations similarity scores for the three types of locations (places, periods, and themes), as defined below.

$$\begin{gathered}L_{type}(E_{ref}, E_{target}) = \\ \frac{\sum_{location \in \text{getLoc}(E_{ref}, type) \cap \text{getLoc}(E_{target}, type)} w^{ref}_{freshness}(location.getPost()) + w^{target}_{freshness}(location.getPost())}{\sum_{location \in \text{getLoc}(E_{ref}, type) \cup \text{getLoc}(E_{target}, type)} w^{ref}_{freshness}(location.getPost()) + w^{target}_{freshness}(location.getPost())} \\ \text{with } type \in \{spatial, temporal, thematic\}\end{gathered} \tag{4.22}$$

- $type$ is the type of locations considered (*spatial*, *temporal*, or *thematic*).
- $location$ is a location (as a reminder, we consider three types of locations: places, periods, and themes).
- getLoc($E_{ref}, type$) returns the set of locations mentioned by E_{ref} of type $type$ (for example, all themes or places mentioned by the user or group E_{ref}). The same applies to E_{target}.
- getLoc($E_{ref}, type$) $\cap$ getLoc($E_{target}, type$) represents locations that are mentioned by **either** E_{ref} or E_{target} (e.g., themes, places, or periods mentioned by E_{ref} or E_{target}). .
- getLoc($E_{ref}, type$) $\cup$ getLoc($E_{target}, type$) represents locations that are mentioned by **both** E_{ref} and E_{target} (e.g., themes, places, or periods mentioned by both E_{ref} and E_{target}).
- $w^{ref}_{freshness}(location.getPost())$ represents the time decay factor $w_{freshness}$ applied to the most recent $post$ where $location$ is mentioned by E_{ref}. Similarly, $w^{target}_{freshness}(location.getPost())$ applies to E_{target}.

This formula allows us to assess whether users or groups are similar based on: (1) where they interacted (spatial, *where*), (2) when they were active (temporal, *when*), and (3) what they interacted about, their interests (thematic, *what*).

Pattern 2 and 3

When applying *Pattern 2* and *Pattern 3* (e.g., user to themes, user to places, etc.), we use the ratio of the users' posts containing the theme to all their posts. This allows us to detect the affinity of users or groups to specific themes, places, and periods. We also weigh occurrences with the time decay factor to give more weight to recent ones. We define $L_{occurrences}(E_{ref}, E_{target})$ as:

$$L_{occurrences}(E_{ref}, E_{target}) = \frac{\sum_{post \in E_{\text{ref}} | E_{\text{target}} \in post.locations} w_{freshness}(post)}{\sum_{post \in E_{\text{ref}}} w_{freshness}(post)}$$
$$\text{with } E_{ref} \in \{user, group\} \text{ and } E_{target} \in \{place, time, theme\} \tag{4.23}$$

- $post \in E_{\text{ref}} | E_{\text{target}} \in post.locations$ represents the set of all posts from user or group E_{ref} that contain E_{target}.
- $\sum_{post \in E_{\text{ref}} | E_{\text{target}} \in post.locations} w_{freshness}(post)$ is the sum of the weights of all posts from user or group E_{ref} that contain E_{target}.
- $\sum_{post \in E_{\text{ref}}} w_{freshness}(post)$ is the sum of the weights of all posts from user or group E_{ref}.

Pattern 4

Lastly, when dealing with *Pattern 4*, which consists in determining the proxemic similarity between places, periods, and themes (e.g., place to themes, theme to periods, etc.), we consider the co-occurrences of *locations* (e.g., places, periods and themes) found within posts. For instance, consider the post: "*We went **swimming** at the **beach** in **Paris***". In this example, "swimming" (*theme*), "beach" (*theme*), and "Paris" (*place*) are co-occurring entities. For this, we can use a variant of the time-weighted locations similarity ($L_{individual}$) score described above (the time decay factor $w_{freshness}$ is the same as defined before).

$$L_{co-occurrences}(E_{ref}, E_{target}) = \frac{\sum_{post \in \text{getPosts}(E_{ref}) \cap \text{getPosts}(E_{target})} w_{freshness}(post)}{\frac{1}{\zeta} \sum_{post \in \text{getPosts}(E_{ref}) \cup \text{getPosts}(E_{target})} w_{freshness}(post)}$$
$$\text{with } E_{ref} \in \{place, time, theme\} \text{ and } E_{target} \in \{place, time, theme\} \tag{4.24}$$

- getPosts(E_{ref}) $\cap$ getPosts(E_{target}) is the set of posts that contain **both** E_{ref} and E_{target}.
- getPosts(E_{ref}) $\cup$ getPosts(E_{target}) is the set of posts that contain **either** E_{ref} or E_{target}.
- ζ is a scaling parameter. It decreases the weight of the denominator, which influences the overall value of the measure. The value to be set depends on the size of the dataset. For now, we default it to 30.

4.6.7 Movement Similarity: $M(E_{ref}, E_{target})$

The *Movement* dimension $M(E_{ref}, E_{target})$ focuses on sequential relationships within multidimensional social media trajectories (note that for now, we do not consider the time intervals in sequences). A sequence is formed when two locations are mentioned consecutively on different posts. Using an example, if a post reads "*We are **visiting** the **museum** in **Paris***", following another "We went *swimming*", sequences like $swimming \rightarrow visiting$, $swimming \rightarrow museum$, and $swimming \rightarrow Paris$ emerge. Here, for simplicity reasons, we limit ourselves to sequences of two contiguous locations of the same nature, representing transitions from one place, period, or theme to another.

Pattern 1

For *Pattern 1* (such as user-to-users or group-to-users similarity), we employ the same formula as location similarity ($L_{\text{individual}}$). However, instead of considering occurrences of *locations* (e.g., places, periods, and themes), we compare the sequencings of these elements across posts. This approach ($M_{individual}$) helps identify common travel patterns among visitors, for example. An illustrative example of the sequencing sets compared is presented in Figure 4.9.

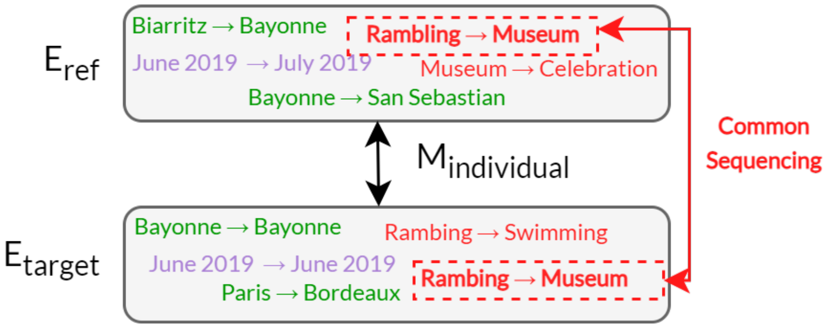


Figure 4.9: Two Sets of Spatial, Temporal, and Thematic Sequencings

Pattern 2 and 3

For *Pattern 2* and *Pattern 3*, the goal is to determine if users consistently stay (or *focus*) in a particular location (e.g., a place, a theme, a period) in their posts. We achieve this by calculating the entropy associated with that location. In the following example, E_{ref} represents a user, and E_{target} a theme.

$$\begin{gathered} M_{entropy}(E_{ref}, E_{target}) = 1 - \left[\left(\frac{C_{\text{target}}}{C_{\text{total}}}\right) \log_2 \left(\frac{C_{\text{target}}}{C_{\text{total}}}\right) + \left(\frac{C_{\text{others}}}{C_{\text{total}}}\right) \log_2 \left(\frac{C_{\text{others}}}{C_{\text{total}}}\right) \right] \\ \text{with } C_{\text{target}} = |\{post \in E_{\text{ref}} \mid E_{\text{target}} \in post.locations\}| \\ \text{and } C_{\text{others}} = |\{post \in E_{\text{ref}} \mid E_{\text{target}} \notin post.locations\}| \\ C_{\text{total}} = C_{\text{target}} + C_{\text{others}} \\ E_{ref} \in \{user, group\} \text{ and } E_{target} \in \{place, time, theme\} \end{gathered} \tag{4.25}$$

1. We start by counting the number of posts containing occurrences of the location E_{target} (this count is denoted as C_{target}). For example, all posts containing a specific place, theme or period. Then, we count the number of posts containing other locations but not E_{target} in E_{ref}'s set of posts (this count is denoted as C_{others}). For example, all posts not containing a specific place, theme, or period.

2. We calculate the total number of posts ($C_{\text{total}} = C_{\text{target}} + C_{\text{others}}$)

3. We compute the entropy ($M_{entropy}$) using the preceding formula (the lower this value is, the less predictable the user's sequence of posts is regarding the reference entity):

Pattern 4

In *Pattern 4*, where we evaluate the relationships among themes, places, or periods (or any combination thereof), we use the conditional probability to determine how frequently one entity consistently appears after another. Given a sequencing set (see Figure 4.9), the probability that an entity E_{target} follows E_{ref} is:

$$M_{sequencing}(E_{ref}, E_{target}) = \frac{N(E_{ref} \cap E_{target})}{N(E_{ref})} \tag{4.26}$$

$$\text{with } E_{ref} \in \{place, time, theme\} \text{ and } E_{target} \in \{place, time, theme\}$$

- $M_{sequencing}$ is the probability that E_{target} occurs after E_{ref}.
- $N(E_{ref} \cap E_{target})$ is the count of times both E_{ref} and E_{target} appear sequentially in users' trajectories.
- $N(E_{ref})$ counts occurrences of E_{ref}.

This measure can help in determining social media users' subsequent destinations after visiting a particular municipality or what people tend to do after practicing a given activity (*theme*). It can also easily be reversed to determine what people do *before*.

4.6.8 Orientation Similarity: $O(E_{ref}, E_{target})$

In the *Orientation* dimension $O(E_{ref}, E_{target})$, both contextual sentiment and engagement data are taken into account, and they are linked on a per-post basis. In our model, sentiment is categorized with labels: *positive*, *negative*, or *neutral*. Engagement is quantified by aggregating metrics such as the number of likes, reposts, comments, etc.

For every proxemic similarity pattern, our formula mirrors that of the *Location* dimensions with one key modification: the weighting. Rather than applying the time-decay factor $w_{freshness}$, we introduce a new orientation factor, $w_{orientation}$, that adjusts post weights based on sentiment and engagement levels associated with it.

$$w_{orientation}(post) = \mu \times post.engagement \times \nu \times post.sentiment \tag{4.27}$$

$$\text{with } engagement \in \mathbb{N} \text{ and } sentiment \in \{0, 1, 2\}$$

- $post.sentiment$ represents the sentiment value of the post $post$, which we map to the following values: $\{0, 1, 2\}$. Here, 0 corresponds to negative posts (*reducing the weight*), 1 to neutral posts (*no change in weight*), and 2 to positive posts (*increasing the weight*). This scaling, for instance, helps in identifying whether multiple users or groups favor the same content, or in emphasizing positive posts when establishing connections between themes or places (as positive posts will weight more than neutral ones and negative posts will not be considered).
- $post.engagement$ stands for the engagement value of the post $post$. A post with higher engagement will carry greater weight. For example, this can help to discern whether various users gain popularity around similar themes. Lastly, μ and ν are coefficients employed to prioritize either sentiment or engagement in the weighting calculation.

Formulas with $w_{orientation}$ weighting are called $O_{individual}$, $O_{occurrences}$ and $O_{co-occurrences}$. We will not define them here, as they are essentially the same formula as above, but with a different weighting.

4.6.9 Implementation of the *ProxMetrics* Toolkit

It is important to note that the *ProxMetrics* toolkit is extensible. While we have chosen to implement it using these formulas, users are free to select others, provided they adhere to the proxemic patterns and the APs Trajectory Model. Table 4.4 offers a summary of the appropriate formulas for each pattern.

	E_{ref}	E_{target}	**D**	**I**	**L**	**M**	**O**
Pattern 1	👤 ∨ 👥	👤 ∨ 👥	$D_{physical}$	$I_{individual}$ I_{group}	$L_{individual}$	$M_{individual}$	$O_{individual}$
Pattern 2	👤 ∨ 👥	📍 ∨ 🕒 ∨ 📖	*n/a*	I_{group}	$L_{occurrences}$	$M_{entropy}$	$O_{occurrences}$
Pattern 3	📍 ∨ 🕒 ∨ 📖	👤 ∨ 👥	*n/a*	I_{group}	$L_{occurrences}$	$M_{entropy}$	$O_{occurrences}$
Pattern 4	📍 ∨ 🕒 ∨ 📖	📍 ∨ 🕒 ∨ 📖	$D_{physical}$ $D_{semantic}$ $D_{interval}$	I_{group}	$L_{co-occurrences}$	$M_{sequencing}$	$O_{co-occurrences}$

Table 4.4: Implementation of the *ProxMetrics* Toolkit. Each line Represents a Proxemic Similarity Pattern from Figure 4.7

We will now experiment with the toolkit, adjusting the dimensions to meet domain-specific requirements, and then qualitatively evaluate the proxemic patterns.

4.7 Experimentation of *ProxMetrics* on the Domain of Tourism

Firstly (Subsection 4.7.1), we demonstrate how tourism offices' requirements can be expressed as proxemic similarity indicators using the *ProxMetrics* toolkit (e.g., choice of proxemic pattern, relevant dimensions). We then elaborate on a chosen request from tourism offices (Subsection 4.7.2) and provide an overview of other selected indicators (Subsection 4.7.3). Lastly, we generalize from this and qualitatively evaluate each proxemic pattern (Subsection 4.7.4). Note that a separate experiment to demonstrate the toolkit's applicability to another application domain (*local public policies*) and requirements will be conducted later, in a dedicated chapter (Chapter 6).

4.7.1 Designing Tourism Indicators with *ProxMetrics*

For this experimentation, we leverage the dataset of 2,961 tweets from visitors in the French Basque Coast Area we have collected in Chapter 2 and then processed in Chapter 3. Figure 4.10 shows a selected extract of the instantiated APs Trajectory Model (note that users were anonymized). Appendix C shows the model extract implemented in JSON format. This JSON structure is used to represent the model in the APs Framework's implementation.

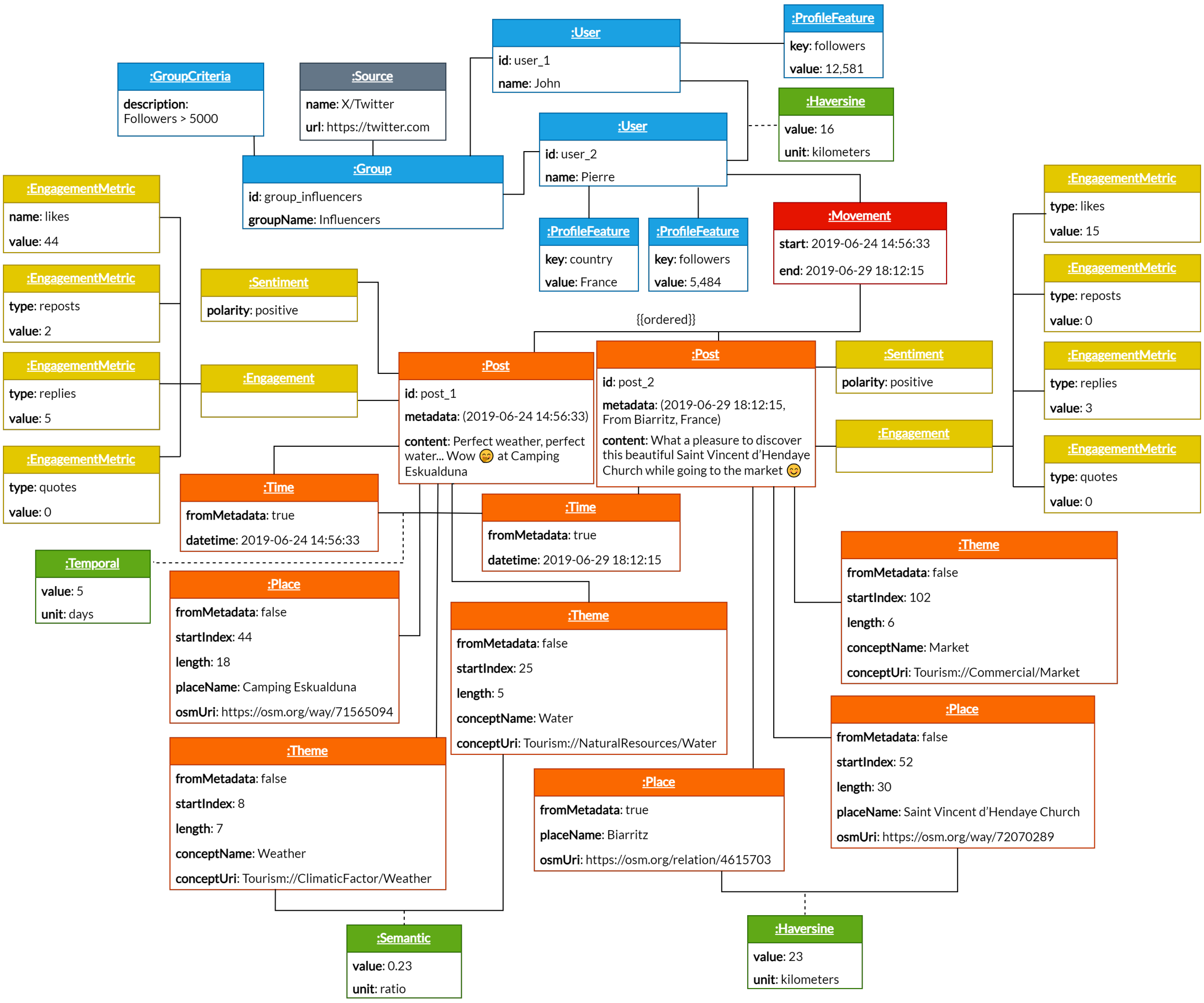


Figure 4.10: Object Diagram of a Subset of the APs Trajectory Model, Instantiated with Touristic Tweets from the Basque Country

Additionally, we have collected requirements from local tourism stakeholders in the *Basque Country* region, specifically the *Tourism Office of the Basque Country*[3] (refer to Table 1.1). Below is an extract of requirements for indicators in the tourism domain, compiled through collaborative discussions with these local tourism stakeholders on the *French Basque Coast*. This tourism office is looking for indicators on:

1. *Leisure Activities Practiced Together* (which activities do visitors often practice together)
2. *Municipalities Visited in Sequence* (after visiting a given municipality, where do visitors usually go)
3. *Municipality-Specific Demographics* (categories of visitors per municipality)
4. *Weather-Based Municipality Preferences* (choice of municipality by visitors based on the weather)

[3] https://www.en-pays-basque.fr

5. *Seasonal Activity Preferences* (choice of leisure activity by visitors based on the season)
6. *Identification of Popular Events* (which events are popular in the region)
7. *Lacking Tourism Infrastructure* (places where infrastructure is found disappointing)
8. *Visitors Satisfaction about POIs* (what POIs do visitors primarily enjoy)
9. *Trends in Cross-Border Tourism* (who, where, when, and what)
10. *Connection of Similar Visitors* (for a visitors' connection system)

These requirements vary in scope, with some being broad and others more specific. Tourism stakeholders require indicators to better understand these diverse aspects of tourism in their region. To address this, we will take advantage of the *ProxMetrics* toolkit to address these requirements and calculate relevant proxemic similarity indicators, allowing for a deeper understanding of the various aspects of tourism in the region.

Table 4.5 presents the requirements from tourism offices as shown above. Each line number corresponds to a requirement from the list. For each requirement, we have proposed a manner to express it as a proxemic similarity indicator using the *ProxMetrics* toolkit. This includes (1) the proxemic environment that could be used to model the requirement, identifying both the reference entity (E_{ref}) and the target entities (τ), whose proxemic similarities to the reference element will be assessed; and (2) the relevant proxemic dimensions of our model that are essential for addressing it. While a variety of dimension combinations and proxemic environments (e.g., reference and target entities) may be applicable for most requirements, we present only a selected one here due to space constraints.

	Proxemic Environment		**Dimensions**				
Req.	**Reference (E_{ref})**	**Targets (τ)**	**D**	**I**	**L**	**M**	**O**
1	📖 Leisure Activity	📖 Leisure Activities			•		•
2	📍 Municipality *or* 📍 POI	📍 Municipalities, POIs	•			•	
3	👥 User Group	📍 Municipalities		•	•		•
4	📖 Weather	📍 Municipalities			•		
5	📖 Leisure Activity	🕒 Seasons	•		•		
6	🕒 Season *or* 📍 Municipality	📖 Events	•		•		•
7	📍 Municipality	📖 Infrastructures			•		
8	📍 Municipality	📍 Points of Interest			•		•
9	📖 FrontierArea	👥 Groups, 📍 Municipalities, 🕒 Periods		•	•	•	•
10	👤 User	👤 Users	•	•	•	•	•

Table 4.5: Example of End-Users Requirements for Social Media Analysis: The Case of Tourism

Let's examine the requirement "*Municipality-Specific Demographics*" (requirement 3 in Table 4.5). In this context, the goal is to calculate indicators allowing tourism stakeholders to identify the types of visitors most commonly associated with various municipalities. Using the *ProxMetrics* toolkit, this requirement can be modeled by selecting a specific user group (*e.g., short-stay visitors*) as the proxemic reference and then determining the proxemic similarity of different municipalities to this group. Therefore, it falls into the proxemic *Pattern 2* (refer to Table 4.4), which links a dynamic reference entity (in this case, a user group) to a static entity (in this case, a place).

To calculate this similarity, it may be useful to take into account multiple proxemic dimensions to achieve a more comprehensive analysis, such as the characteristics of the reference user group and users visiting municipalities (*Identity* dimension), the frequency with which members of the reference group mention the municipality in their posts (*Location* dimension), and the sentiments they generally express towards it (*Orientation* dimension). For example, a given group of visitors and municipality could be perceived *similar* if positive sentiments are frequently expressed by the group about the municipality. This requirement could be articulated in various alternative ways (e.g., by choosing a specific municipality as the reference point instead of a user group). As the toolkit is modular, the end-user is able to give more importance to certain dimensions by adjusting the dimensional weighting. In summary, Table 4.5 shows that the *ProxMetrics* toolkit is versatile and capable of modeling various requirements. We will now go into details on a selected one.

4.7.2 Indicator Case Study: Municipality-Specific Demographics

We now go into more detail with the example requirement presented above. Figure 4.11 focuses on the indicator "*Municipality-Specific Demographics*".

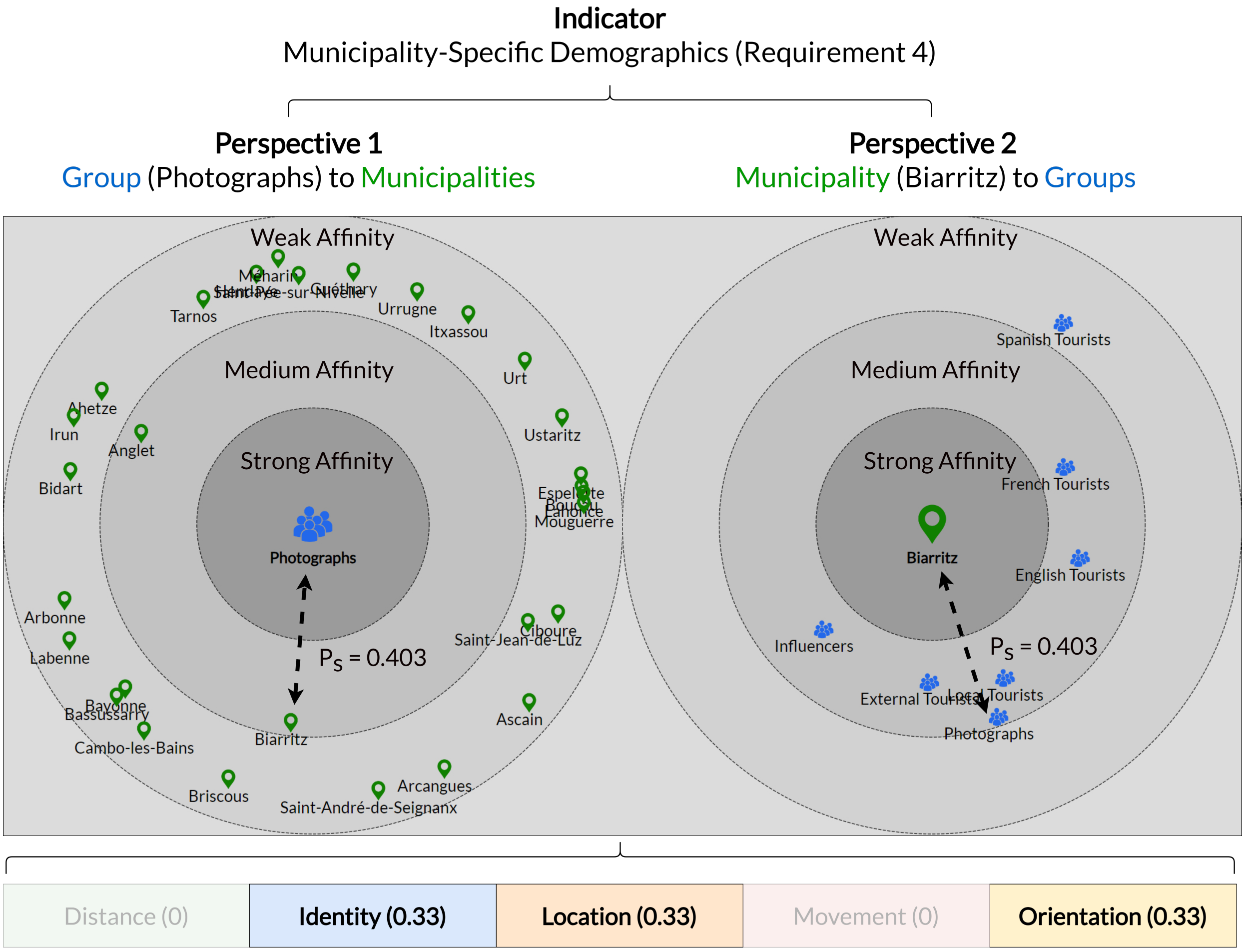


Figure 4.11: Visualization of the Indicator "*Municipality-Specific Demographics*" in Proxemic Reticles, Using Proxemic Patterns 2 and 3

Here, *proxemics* is used for two purposes: presenting proxemic similarity as a distance between social media entities, focusing on a reference one, and computing this similarity using a blend of proxemic dimensions that are adapted to the application domain and requirement of interest.

Visuals were created using a dedicated interface powered by the *ProxMetrics* toolkit, able to visualize results through proxemic reticles. This interface also helps end-users set up the toolkit (dimensions, entities, etc.). We will present it in Chapter 5.

In this example, the *"Municipality-Specific Demographics"* indicator is expressed using two distinct proxemic environments (*Perspective 1* and *2* in Figure 4.11). One where the demographic (user group entity) is the reference (here, *Photographers*), and *close* municipalities are positioned relative to it, and another perspective where a chosen municipality of interest (here, the touristic municipality of *Biarritz*) is at the center, and we want to identify groups *close* to this municipality. These two perspectives correspond to the proxemic *Pattern 2* and *Pattern 3* defined earlier (see Figure 4.7). By default, we set proxemic dimensions (ILO) to equal weighting ($\frac{1}{3} \approx 0.33$), but this default weighting can be dynamically changed by end-users if they wish to give more weight to profile features or municipality mentions.

- The *Identity* (I) dimension influences the result based on whether the typical user profiles of a municipality are similar to the group based on profile features (e.g., are the visitors of the municipality of *Biarritz* usually of the same age compared to photograph visitors, etc.).
- The *Location* (L) dimension is based on how often members of the user groups mention the given municipality in their posts, which may indicate a specific affinity for the municipality.
- The *Orientation* (O) dimension is considered and weights municipality mentions by sentiment and engagement values, which means that influential or positive users will have more impact on the similarity results.

Let's illustrate *Perspective 1* from Figure 4.11 using the pairing $E_{ref} = Photographers$ and $E_{target} = Biarritz$. We can refer back to Table 4.4 to determine the formula for this proxemic environment pattern, specifically *Pattern 2*. We calculate the proxemic similarity between *Biarritz* and Photographers as follows, using the I_{group}, $L_{occurrences}$, and $O_{occurrences}$ formula corresponding to the Identity, Location, and Orientation (ILO) dimensions considered in this example. We obtain a proxemic similarity of around 0.403 by aggregating the I, L, and O dimensions with equal weighting. It appears that users classified as *Photographers* have a moderate interest in this municipality when considering these 3 dimensions.

$$\begin{gathered}
P_s(Photographers, Biarritz) = \\
\frac{1}{3} \times \underbrace{I_{group}(Photographers, \text{getUsersMentioning}(Biarritz))}_{0.52} + \\
\frac{1}{3} \times \underbrace{L_{occurrences}(Photographers, Biarritz)}_{0.34} + \\
\frac{1}{3} \times \underbrace{O_{occurrences}(Photographers, Biarritz)}_{0.36} \\
\approx 0.403
\end{gathered} \tag{4.28}$$

In Figure 4.11, results are displayed through a proxemic reticle with the reference entity in the center and the target entities around it, scattered in different proxemic zones. Here, we have 3 zones: *strong affinity*, *medium affinity*, and *weak affinity*. However, other zones could be defined according to the domain and the requirement of interest. Any number of zones is possible. As we can observe, these visualizations are useful and flexible for domain stakeholders as they allow immediate identification of similar or dissimilar entities.

To further elaborate and demonstrate the *ProxMetrics* toolkit's capability to calculate a wide array of indicators, we will provide an overview of four additional selected indicators.

4.7.3 Overview of Additional Indicators

Figure 4.12 presents example results for four additional indicators from those introduced in Subsection 4.7.1. These four indicators were chosen to demonstrate the capabilities of the *ProxMetrics* toolkit, as they leverage various types of entity combinations and proxemic dimensions. The proxemic similarity indicators displayed are as follows:

- *Leisure Activities Practiced Together* (Figure 4.12, *Indicator A*): for this indicator, we are examining similar touristic activity-related themes compared to a reference one, in this example, *Surfing*. We use two equally weighted dimensions: *Location* (L), to consider the co-occurrences of themes in posts, and *Orientation* (O), so that positive co-occurrences weigh more heavily, as we consider those to be more significant. The results show that *Surfing* is often paired with *Photography* (see ①) and *Outings* (see ②). Perhaps, it appears that photographing surfers is particularly popular in the region.

- *Municipalities Visited in Sequence* (Figure 4.12, *Indicator B*): here, we are interested in the spatial movements of visitors, namely, where they tend to go after visiting a given place. In this case, we are focusing on the municipality of *Ustaritz*, positioned as the reference entity, and it will be compared to other municipalities in the region. We use the *Movement* (M) dimension with stronger weighting because we are particularly interested in the places that are often sequenced in visitors' trajectories. Minor weighting is given to the *Distance* (D) dimension to slightly boost places that are physically close, as they are more likely to attract visitors. The results show that the municipalities of *Biarritz* (see ①) and *Saint-Jean-De-Luz* (see ②) are more likely to attract visitors after having explored *Ustaritz*. An extension of this indicator could be to calculate it for specific user categories, for example, *Photographers*. This would allow a recommender system to determine where to recommend another photographer to go based on the behavior of other photographers (via a collaborative filtering approach (Liu et al., 2014)). If *Ustaritz* was not found in any tweets, then only physically close municipalities would be considered in the proxemic reticle.

- *Lacking Tourism Infrastructure* (Figure 4.12, *Indicator C*): for this requirement, the objective is to identify themes related to infrastructure that are broadly considered lacking in a given municipality. We use the municipality of *Biarritz* as a reference and observe similar themes based on the *Orientation* dimension. This dimension will bring positive themes closer to the reference municipality and push negative ones further. The results show that *Shopping* (see ①) and *Exhibitions* (see ②) are quite dissimilar, indicating that these aspects are severely

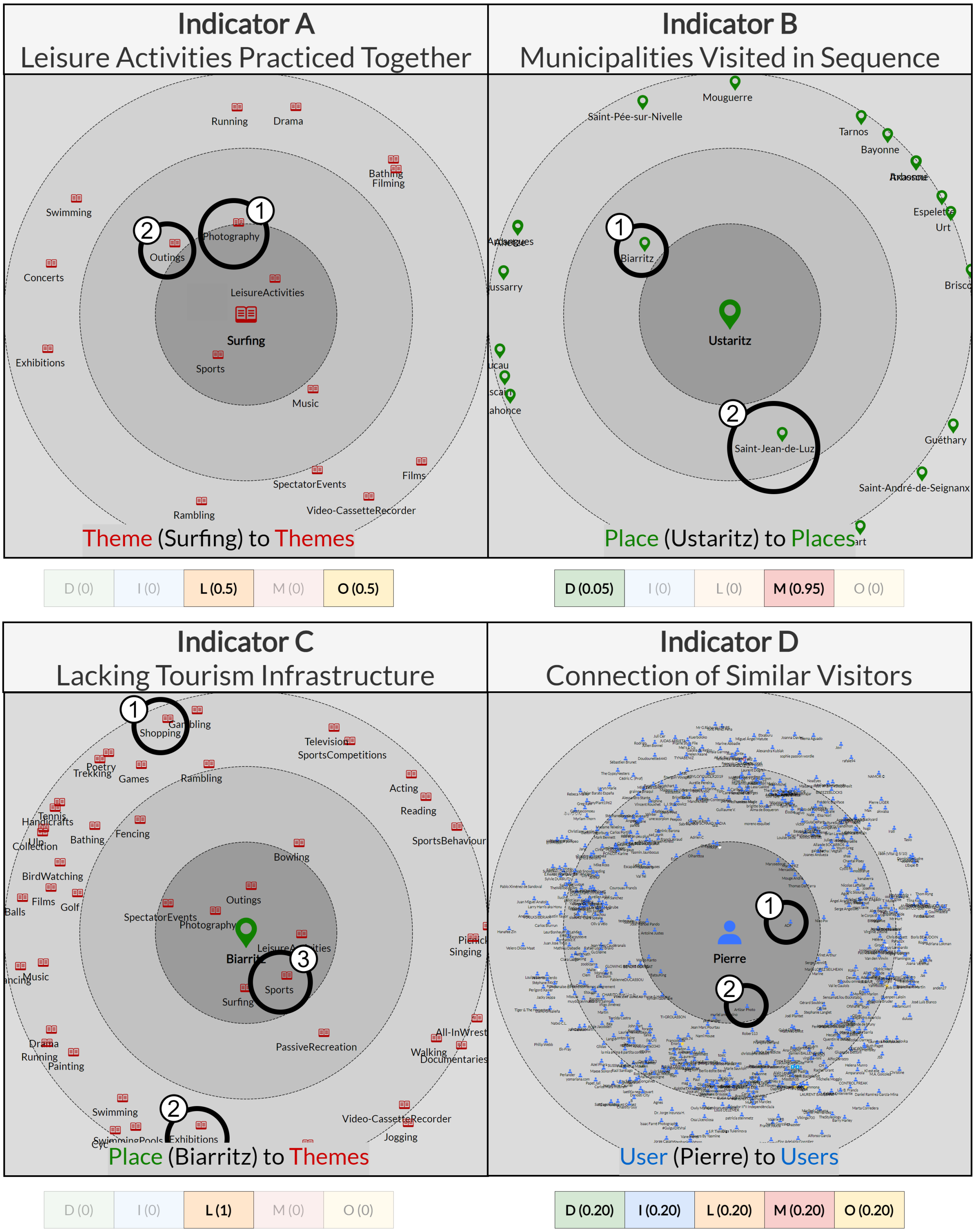


Figure 4.12: Results of Four Selected Proxemic Similarity Indicators from the Tourism Office Requirements

lacking in this municipality. On the other hand, *Sports* (see ③) is very similar, suggesting that the municipality is viewed very positively in this aspect by visitors.

- *Connection of Similar Visitors* (Figure 4.12, *Indicator D*): Finally, the last requirement is to detect similar visitors to build a system that connects them. If we select a reference visitor, here *Pierre*, we can observe close matches. This indicator is based on all dimensions because we want to get the overall similarity based on various criteria. This can be tweaked depending on the use case, as the toolkit is modular. ① and ② show examples of other visitors that are closer to the reference one and could therefore be recommended because they have similar behavior and likely share interests with the reference user. This indicator could also be used for a recommender system to suggest themes or places based on what similar users liked.

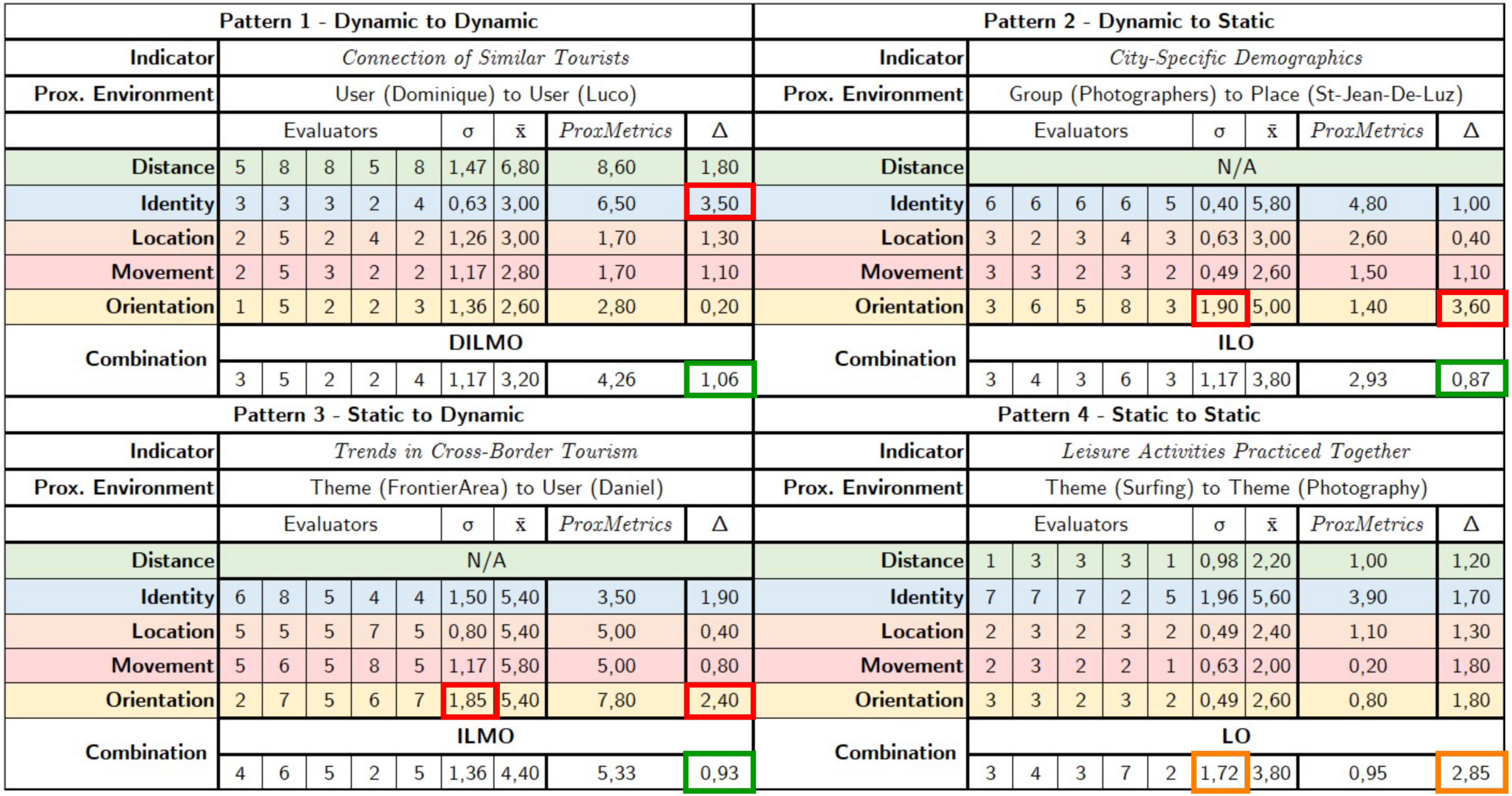
We have demonstrated that *ProxMetrics* effectively models a wide range of diverse indicators in the tourism domain by combining various entities and modular proxemic dimensions according to the requirements of domain stakeholders. However, it is now essential to evaluate the relevance and significance of the produced proxemic similarity indicators for the stakeholders.

4.7.4 Qualitative Evaluation of Proxemic Patterns

We generalize from this and qualitatively evaluate the four defined proxemic patterns with various dimensions, as well as the methodology used to combine them, to assess the validity of our working hypotheses and the relevance of the indicators produced by the toolkit. Additionally, we conclude the experiment by comparing the toolkit with existing platforms dedicated to analyzing touristic data to highlight its complementarity in this domain.

To determine whether the *ProxMetrics* toolkit produces indicators that are accurately representative of real-world phenomena or behaviors on social media, we conducted a qualitative evaluation of the results and compared the results obtained by *ProxMetrics* with the assessments of domain experts (colleagues specialized on cultural heritage and tourism practices), as depicted in Table 4.6. For each identified proxemic similarity pattern (from Figure 4.7), we selected an indicator within this pattern as a case study. As mentioned earlier, the proxemic similarity is typically calculated between a reference entity and multiple target entities. However, for this evaluation, we focused on calculating the similarity between the reference entity and a single target entity to facilitate the work of the experts.

We implemented the following protocol. Firstly, for each of the four proxemic patterns, we asked five domain experts to assess the similarity (on a scale from 1 to 10, with 1 being extremely dissimilar and 10 being extremely similar) of the reference and the selected target entity for each dimension individually. To do this, experts were provided with only excerpts of the corresponding to the given dimension. For example, in the case of *Pattern 1*, when evaluating the proxemic similarity between two users within the *Identity* dimension, two sets of user profiles along with their characteristics (age, number of followers, etc.) were given but no information on the tweets they posted, places visited, etc. When dealing with the *Location* dimension, they were only given two sets of tweets, but not the sequence in which they were issued or the users to whom they belong, etc. This allows us to determine whether our individual dimension formulas are relevant and meaningful to domain stakeholders. Then, we asked the domain experts to assess the similarity of all relevant dimensions combined in regard to the example indicator chosen.



Pattern 1 - Dynamic to Dynamic										Pattern 2 - Dynamic to Static									
Indicator	*Connection of Similar Tourists*									Indicator	*City-Specific Demographics*								
Prox. Environment	User (Dominique) to User (Luco)									Prox. Environment	Group (Photographers) to Place (St-Jean-De-Luz)								
	Evaluators					σ	x̄	*ProxMetrics*	Δ		Evaluators					σ	x̄	*ProxMetrics*	Δ
Distance	5	8	8	5	8	1,47	6,80	8,60	1,80	Distance	N/A								
Identity	3	3	3	2	4	0,63	3,00	6,50	3,50	Identity	6	6	6	6	5	0,40	5,80	4,80	1,00
Location	2	5	2	4	2	1,26	3,00	1,70	1,30	Location	3	2	3	4	3	0,63	3,00	2,60	0,40
Movement	2	5	3	2	2	1,17	2,80	1,70	1,10	Movement	3	3	2	3	2	0,49	2,60	1,50	1,10
Orientation	1	5	2	2	3	1,36	2,60	2,80	0,20	Orientation	3	6	5	8	3	1,90	5,00	1,40	3,60
Combination	DILMO									Combination	ILO								
	3	5	2	2	4	1,17	3,20	4,26	1,06		3	4	3	6	3	1,17	3,80	2,93	0,87

Pattern 3 - Static to Dynamic										Pattern 4 - Static to Static									
Indicator	*Trends in Cross-Border Tourism*									Indicator	*Leisure Activities Practiced Together*								
Prox. Environment	Theme (FrontierArea) to User (Daniel)									Prox. Environment	Theme (Surfing) to Theme (Photography)								
	Evaluators					σ	x̄	*ProxMetrics*	Δ		Evaluators					σ	x̄	*ProxMetrics*	Δ
Distance	N/A									Distance	1	3	3	3	1	0,98	2,20	1,00	1,20
Identity	6	8	5	4	4	1,50	5,40	3,50	1,90	Identity	7	7	7	2	5	1,96	5,60	3,90	1,70
Location	5	5	5	7	5	0,80	5,40	5,00	0,40	Location	2	3	2	3	2	0,49	2,40	1,10	1,30
Movement	5	6	5	8	5	1,17	5,80	5,00	0,80	Movement	2	3	2	2	1	0,63	2,00	0,20	1,80
Orientation	2	7	5	6	7	1,85	5,40	7,80	2,40	Orientation	3	3	2	3	2	0,49	2,60	0,80	1,80
Combination	ILMO									Combination	LO								
	4	6	5	2	5	1,36	4,40	5,33	0,93		3	4	3	7	2	1,72	3,80	0,95	2,85

Table 4.6: Qualitative Evaluation of the *ProxMetrics* Toolkit on the Four Proxemic Similarity Patterns (See Figure 4.7) with Selected Indicators.

This helps us to determine whether our method of combining dimensions (weighted mean) is relevant. For this evaluation, we deliberately selected entities that are not overly represented in the entire dataset, aiming to facilitate the experts' evaluation process.

As seen in Table 4.6, for each dimension and pattern, we calculated the standard deviation (depicted as σ) between the five experts' measures (1, 2, 3, 4, and 5) in order to assess their degree of agreement. Then, we averaged the measures provided by the experts (depicted as $\bar{x}$) and compared them against the results obtained by *ProxMetrics*. This allowed us to calculate the difference (depicted as Δ) between the experts' results and those of *ProxMetrics*, thus determining whether they are in concordance or not.

As we can observe in Table 4.6, for the individual dimensional evaluation, out of the 18 evaluation cases, there are only 3 cases (🟥 in Table 4.6) where the *ProxMetrics* assessment significantly differs from that of the evaluators (e.g., the Δ is greater than 2). These are *Pattern 1* with the I dimension ($\Delta = 3.50$), *Pattern 2* with the O dimension ($\Delta = 3.60$), and *Pattern 3* with the O dimension ($\Delta = 2.40$). In cases of poor assessment for the O dimension, we notice they are also correlated with very low agreement between evaluators ($\sigma = 1.90$ for *Pattern 2* and $\sigma = 1.85$ for *Pattern 3*), indicating there are various ways to interpret the similarity for this dimension. Therefore, other formulas might be more appropriate depending on the domain requirements. For the other dimensional measures, *ProxMetrics* provides assessments that are quite similar to those of the experts, demonstrating the relevance of our formula in the domain of tourism.

In the case of multidimensional combinations (DILMO, ILO, ILMO, and LO in the examples chosen), for most of them (3 out of 4, 🟩 in Table 4.6), the aggregation with default parameters (equal weighting) of *ProxMetrics* closely matches that of the evaluators, with $\Delta = 1.06$, $\Delta = 0.87$, and $\Delta = 0.93$. However, for *Pattern 4*, the result diverges with $\Delta = 2.85$ (🟧 in Table 4.6), and the agreement among evaluators is also significantly weaker ($\sigma = 1.72$) compared to the other patterns. In all cases, it would have been possible to improve the accuracy of the results by altering the weighting of each dimension. Let's consider *Pattern 1* as an example: if we had doubled the impact of the L and O dimensions and reduced that of D by half, we would have obtained a result of ≈ 3.30, similar to that of the experts ($\frac{(8.6\times0.5+6.5+1.7\times2+1.7+2.8\times2)}{6.5}$). This highlights the importance of choosing appropriate weights during the process to get accurate results. These weight values are highly dependent on the domain and specific requirements; therefore, experts in each domain must tweak them.

Let's conclude this experiment by comparing the *ProxMetrics* toolkit applied to the domain of tourism with indicators produced by other platforms in this domain. *ProxMetrics* is highly complementary for several reasons:

- *Dynamic, User-Parameterized Indicators*: It is a dynamic and modular tool that allows users to build their own indicators in real time using proxemic dimensions (DILMO). This contrasts with static indicators in dashboards such as Pilat Tourisme (2022), Isère Attractivité (2023), or Atout France (2023), where users have no say about what is presented to them.

- *Distinctive Analytical Insights*: *ProxMetrics* introduces a type of indicator (*proxemic similarity*) rarely seen in existing solutions. This allows tourism stakeholders to gain new analytical perspectives by assessing the similarity of domain-specific entities (*touristic activities, points of interest, types of visitors, etc.*) based on multiple criteria. Additionally, usual tourism dashboards like the one from Visit Paris Region (2023) usually allow combined filtering but are limited in terms of blending multidimensional indicators together.

- *New Perspective (Proxemic Reticles)*: This offers a mode of visualization that contrasts with the classic visualizations used in dashboards such as UNWTO (2023), INSEE (2023a), or OECD (2023), which contain spatial maps, timelines, tables, and numerical charts but do not highlight the similarity of entities relative to each other.

In Chapter 6, we will demonstrate the toolkit's genericity on other application domains, focusing on the domain of *local public policies*. However, for now, we will focus on avenues to expand this work and tackle its limitations.

4.8 Summary and Perspectives

Here, we proposed a redefinition of the theory of *proxemics* for use in social media. The goal is to provide a solid foundation for designing domain-adaptable indicators for use in social media. Indeed, from our review of the literature, current indicators are either too domain-specific or not comprehensive or modular enough to accommodate a wide range of requirements across various domains. To alleviate this challenge, we proposed a formal redefinition of the theory of *proxemics* (Hall, 1966; Hall et al., 1968), traditionally applied to physical spaces, for application in social media contexts. Building upon our formal redefinition of *proxemics*, we introduce a proxemic-based data model along with *Object Constraint Language* (OCL) constraints specifically designed for modeling social media entities, along with their trajectories and interactions, in a domain-independent manner. This model, called the APs Trajectory Model, is multidimensional, drawing upon the five dimensions of *proxemics* (Greenberg et al., 2011) redefined for use within social media. It is designed to be modular and extensible to accommodate a wide range of use cases and requirements. Unlike existing social media models, it also incorporates the concept of proximity into digital social media spaces. Finally, leveraging this foundation, we designed *ProxMetrics*, a modular and generic toolkit accompanied by formulas to compute domain-adaptable indicators for social media data analysis. This toolkit enables the representation of these indicators as proxemic similarity metrics between multidimensional social media entities, including, but not limited to, users, groups, places, themes, and periods. These indicators are customizable, allowing for the modulation of the five dimensions of *proxemics* to meet various domain-specific requirements.

The formal redefinition of *proxemics* and the associated model were qualitatively evaluated through the analysis of a tweets corpus focused on the *Basque Country* region and the tourism domain. The proxemic trajectory model was instantiated using the most efficient NLP techniques identified in the previous chapter. Subsequently, we gathered requirements from local tourism stakeholders (specifically, the *Tourism Office of the Basque Country*) to employ the toolkit in generating indicators pertinent to their requirements. The outputs of the toolkit were assessed by human evaluators, and in the majority of instances, the results aligned with the assessments of human evaluators (15 out of 18 cases). Another application of these proposals will be shown in Chapter 6 on another domain of application.

We have identified several limitations of this work that we are attempting to mitigate. Below are some future perspectives to improve and extend this work.

Firstly, from the experimental standpoint, we plan to conduct a more extensive evaluation of the toolkit on larger datasets to determine whether it scales effectively to massive volumes of data, on the order of millions of posts, which is not uncommon in social media analyses (Zhao et al., 2012). Assessing the computational cost of the formula used is crucial for achieving real-time results. Additionally, our experiments pointed out that in three cases, the results significantly diverged from human assessments. In such cases, it would be beneficial to investigate alternative formulas for each proxemic dimension. Currently, we have chosen to implement each proxemic dimension based on existing, commonly used formulas that we have sometimes modified and extended. Exploring other formulas to assess whether they perform better is an interesting avenue for future research. For instance, as reflected in Table 4.3, graph-based approaches have proven to be efficient in calculating similarity between social media users (Kumar and Vineela, 2020; Wang et al., 2010b).

Secondly, we envisage extending proxemic applications. For now, the toolkit is limited solely to calculating proxemic similarity indicators for domain stakeholders, but it could be adapted for a variety of purposes beyond this. The initial application could be to leverage the proxemic similarity measurements produced by *ProxMetrics* as inputs for a recommender system. The toolkit is capable of computing similarities among various social media entities, including users, groups, themes, places, or periods, with support for heterogeneous combinations. Consequently, it would be natural to extend the work by proposing a generic recommender system capable of offering recommendations across multiple application domains dynamically in a way that is configurable by end-users, this is an active research topic (Hao et al., 2024; Zang et al., 2022; Zhu et al., 2021). For instance, in the tourism domain, this could involve recommending touristic activities or Points of Interest (POIs) or creating a user recommender system to connect visitors with shared interests. We also envision the integration of proxemic dimensions into a domain-specific language (DSL) specifically designed for querying social media. This DSL would empower users to conduct visual queries of social media entities by manipulating proxemic dimensions. As proxemic dimensions are concrete and directly related to physical space, they offer a valuable tool for individuals lacking a computer science background to formulate queries within social media datasets. We will detail this further in the general perspectives section at the end of the manuscript (see Section 7.2).

Lastly, we envisage using the *ProxMetrics* toolkit for the detection of bots and avatars on social media, addressing one of the world's pressing issues (Ferrara, 2023). The *ProxMetrics* toolkit can calculate proxemic similarities between users, which may be indicative of automated behavior. For effective bot detection, it will be necessary to incorporate additional dimensions, such as temporal activity patterns (Chavoshi et al., 2017), interaction diversity (Kosmajac and Keselj, 2019), linguistic consistency (Cardaioli et al., 2021), and network centrality (Shinan et al., 2023), to enhance the accuracy and reliability of the detection process. We will not develop this perspective as it falls outside the scope of this thesis.

We will now move on to the last phase of the APs Framework (*Valorize* in Figure 1.5), where social media analyses, including the proxemic similarity indicators introduced here, are presented to non-computer scientists end users.

Chapter 5

Valorize
Interactive Visualization of Multidimensional Analyses from Social Media

"Visualization gives you answers to questions you didn't know you had."
— Ben Schneiderman, Computer Scientist

In today's data-driven era, the ability to quickly visualize, and make decisions based on vast amounts of data has become crucial in various domains of application (Sarker, 2021). This is particularly true for social media platforms, which generate extensive amounts of textual data. This chapter, aligned with the *Valorize* phase of the APs Framework (see Figure 1.5), intersects with the research fields of *Human-Computer Interaction (HCI)*, *Interactive Systems and Tools*, and *Visualization*. It addresses the challenge of efficiently presenting multidimensional analyses on social media to non-computer scientist users (e.g., domain stakeholders). Note that while there are various methods to valorize social media analyses, this thesis will focus exclusively on data valorization through visualizations dedicated to end users.

Insights for End-Users
Non-Computer Scientists
Collect
Transform
Analyze
Valorize

We hypothesize that a generic and interactive dashboard, centered around four core dimensions: spatial, temporal, thematic, and personal, augmented with data enrichment elements such as sentiment and engagement metrics, and drawing inspiration as well as blending various design principles of Business Intelligence (BI), Geographic Information Systems (GIS), and Linguistic Information Visualizations, could offer a user-friendly and adaptable platform for non-computer scientists to easily gain insights into social media annotations (introduced in Chapter 3) and derived indicators (e.g., the proxemic similarity indicators introduced in Chapter 4).

The organization of this chapter is as follows. We start by underscoring the significance of visualization-based decision support tools across various domains (Section 5.1) and undertake a review of existing literature across five primary areas: Domain-Specific Dashboards, Geographic

Information Systems (GIS), Business Intelligence Platforms (BI), Linguistic Information Visualizations and Generic Visualization Libraries (Section 5.2). Building on the strengths of these tools, we introduce our contribution: *TextBI* (Masson et al., 2024c), a domain-adaptable dashboard designed to visualize multidimensional social media analyses (Section 5.3, Contribution 4). A demonstration video of the platform is provided at the following address[1]. *TextBI* is then experimented with in the tourism domain (Section 5.4) and qualitatively evaluated through active collaborations with local tourism offices (Section 5.5). Lastly, we propose perspectives to enhance the *TextBI* platform and address its limitations (Section 5.6). The proposal of this chapter has been published in the following journals and venues.

- M. Masson, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose, R. Agerri. (2024). TextBI: An Interactive Dashboard for Visualizing Multidimensional NLP Annotations in Social Media Data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2024)* (pp. 1-9) (St. Julians, Malta). Association for Computational Linguistics. (Core Rank: A, ERA Rank: A)

- M. Masson, C. Sallaberry, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose, R. Agerri. (2024). TextBI: Interactive Visualization of Multidimensional Data from Social Media. In *Mappemonde. Quarterly Journal on the Geographic Image and Forms of the Territory* (to be published)

- M. Masson, S. Abdelhedi, C. Sallaberry, R. Agerri, M. N. Bessagnet, A. Le Parc Lacayrelle, P. Roose. (2023). Interactive Visualization of Touristic Activity Trajectories: Application to Data Extracted from Twitter. In *Workshop "Exploring Traces in an All-Digital World: Challenges and Perspectives" at INFORSID 2023* (La Rochelle, France).

Additionally, the *TextBI* platform won 1st place[2] at the *GeoData Challenge* of the *National Geonumeric Days 2023* (GeoDataDays 2023).

5.1 Introduction: From Multidimensional Text Analyses to Domain-Specific Insights

NLP and Information Extraction (IE) pipelines play a crucial role in transforming unstructured text data, such as social media posts, into structured knowledge (Souili et al., 2015). However, the vast quantity of social media information and the wide range of potential automatic annotations can make it challenging to efficiently extract actionable insights from annotated social media data, for example, to help in decision-making processes in various domains. Consequently, there is a requirement for tools that can facilitate the interpretation and understanding of automatic annotations and indicators within a social media corpus for people who are not necessarily computer scientists.

In the context of the APs Framework, we have two levels of data. The first one is instantiated according to the APs Trajectory Model (refer to Figure 4.5) and is composed of social media posts, their metadata (such as, for example, geotags, number of likes, replies, reposts, etc.), the

[1] https://maxime-masson.github.io/TextBI

[2] https://www.geodatadays.fr/page/GeoDataDays-2023-Les-Challenges-Geodata/139

sequencing between posts, multidimensional automatic annotations of sentiment, places, and thematic concepts, linked with a semantic resource, as well as static distances computed between social media entities. Then, on a second level, we have proxemic similarity measures, which are multi-criteria indicators calculated on top of this model using the *ProxMetrics* toolkit (refer to Figure 4.1). The annotations and indicators in their current raw state present interpretation challenges for the target demographic of our framework, which primarily comprises non-computer scientist users such as domain stakeholders. We are therefore confronted with the challenge of presenting social media analyses (annotations and indicators) to non-computer scientists in a domain-adaptable (any domain of application) and source-independent (any social media) manner, making the findings accessible and understandable to those without a background in computer science. Figure 5.1 shows a simplified view of the APs Framework pipeline and the role of visualization in it.

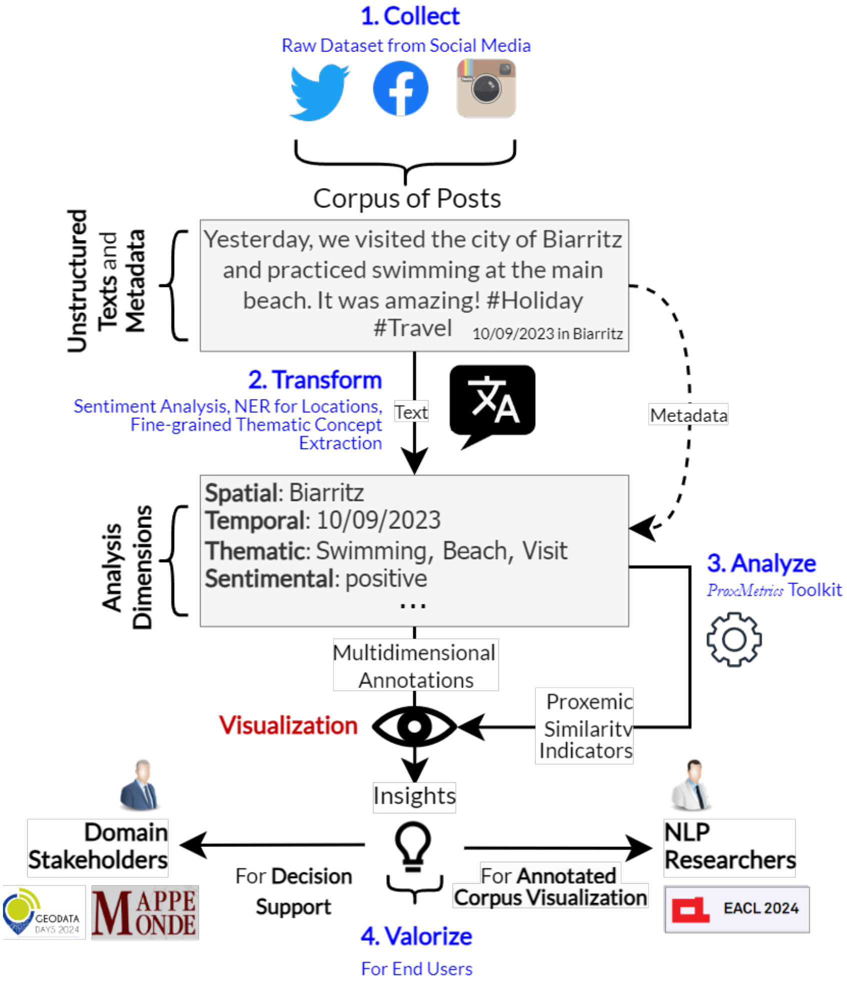

1. Collect
Raw Dataset from Social Media
Corpus of Posts
Unstructured Texts and Metadata
Yesterday, we visited the city of Biarritz and practiced swimming at the main beach. It was amazing! #Holiday #Travel
10/09/2023 in Biarritz
2. Transform
Sentiment Analysis, NER for Locations, Fine-grained Thematic Concept Extraction
Text
Metadata
Analysis Dimensions
Spatial: Biarritz
Temporal: 10/09/2023
Thematic: Swimming, Beach, Visit
Sentimental: positive
...
3. Analyze
ProxMetrics Toolkit
Multidimensional Annotations
Proxemic Similarity Indicators
Visualization
Insights
Domain Stakeholders
GEODATA DAYS 2024
MAPPE MONDE
For Decision Support
For Annotated Corpus Visualization
NLP Researchers
EACL 2024
4. Valorize
For End Users

Figure 5.1: The Role of Visualization in the APs Framework

We aim at proposing a platform (Figure 5.1, *Visualization*) that is designed to address the requirements of two distinct categories of users:

- *Domain Stakeholders* (Figure 5.1 and Figure 5.2) seeking specific insights related to their domain for decision support. This category of users is the main target of the visualization platform. For example, in the tourism industry, tourism offices may find it beneficial to analyze certain types of information. This could include identifying the most popular visitors' activities, determining which municipalities are often visited together, and understanding the emotions or opinions of visitors regarding their experiences. This information can assist them in making informed decisions. As a reminder, our aim is not to fully replace existing visualization and decision support tools in various domains but rather to enrich them with insights from social media. A publication in the *Mappemonde*[3] journal (a journal dedicated to non-computer scientist users, namely geographers) and the *GeoDataDays* prize validates this target.

- *NLP Researchers* (Figure 5.1 and Figure 5.2). We hypothesize that this platform could address some requirements of NLP researchers looking for a tool to visualize annotated social media corpora. This could involve observing the distribution of various types of NLP annotations to observe recurrent themes, places, or sentiments. In the same way as *domain stakeholders*, we do not aim to replace existing tools and methods, but rather to enrich them with a higher-level platform. A publication at the conference of the *European Chapter of the Association for Computational Linguistics* (EACL) (Masson et al., 2024c), a conference dedicated to NLP researchers, validates this target.

Figure 5.2 presents an overview of the requirements we gathered (1) in the context of our project and (2) sourced from both categories of users involved.

1. Some requirements are global to the APs Project and not linked to specific categories of users (Figure 5.2, ①), such as the domain and source independence evoked before, as well as the versatility to adapt to cover various requirements in each application domain. These requirements have been partially addressed in the preceding chapters dedicated to the phases of *Collect* (see Chapter 2), *Transform* (see Chapter 3), and *Analyze* (see Chapter 4). We are now addressing the *Valorize* phase, which we have chosen to undertake through visualization.

2. Others are associated with stakeholders in various domains who are non-computer scientist users (Figure 5.2, ②). For this category of users, from the various low-level requirements presented in Table 1.1 collected from the *Tourism Office of the Basque Country*[4], we have identified and generalized recurrent high-level aspects that domain stakeholders often want to analyze. These aspects include general satisfaction, mentioned places and topics, time, engagement, as well as user and associated profiles. Domain stakeholders want to analyze these aspects through the lens of various indicators: *frequency* (for example, the proportion of domain-specific topics in social media corpora), *association* (for example, whether several places are often associated by social media users), *trajectories* (for example, the movement of a user over time), etc. They also require multi-criteria indicators that combine several of them to address more complex requirements and interactivity, namely, the ability to filter, aggregate, and change the granularity of the social media data displayed, as well as adapt the indicators displayed, all of this in an easy-to-use manner for people with no computing expertise to help in the decision-making processes.

[3] *Quarterly Journal on the Geographic Image and Forms of the Territory*

[4] `https://www.en-pays-basque.fr`

3. Finally, we have requirements of NLP researchers, who are computer scientist users (Figure 5.2, 3). This category of user is **not specifically targeted by our framework**, but by exchanging with them, we noticed that many requirements are quite similar to those of domain stakeholders, but expressed using a different (linguistic-based) vocabulary. For example, NLP researchers observe text and token annotations, such as sentiment, named entities, and thematic concepts, or view their *distribution* in a social media corpus, frequently *co-occurring* or *sequenced* annotations, as well as quantitative statistics about a corpus (e.g., the total number of users, posts, unique concepts, and places, etc.). They also want interactivity (e.g., filtering, aggregation, granularity change) to have different perspectives of the data. The objective is for them to qualitatively evaluate the result of NLP pipelines and models. Furthermore, there is a requirement for extensibility to accommodate the various types of text and token annotations that exist.

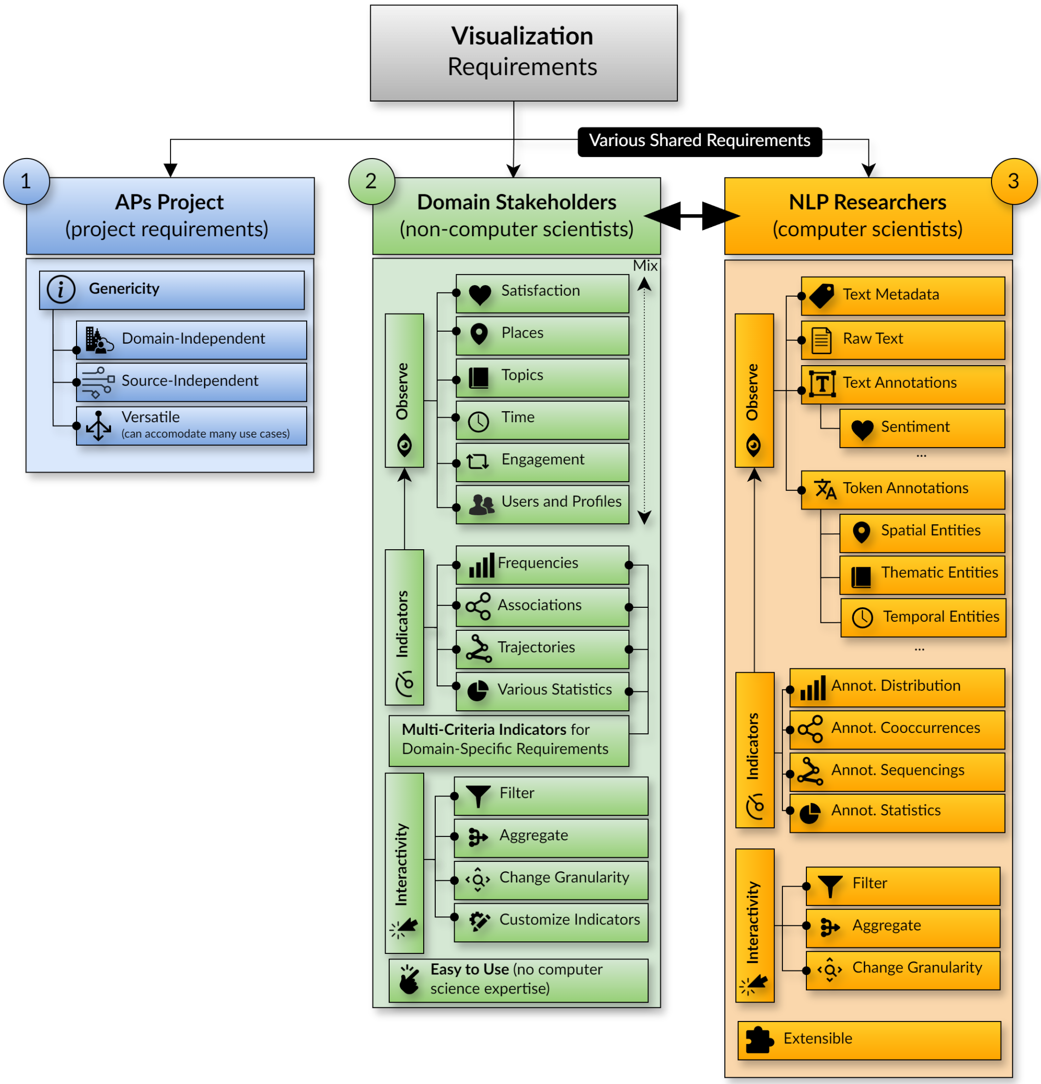


Figure 5.2: Category of End Users and Visualization Requirements

We hypothesize that a generic dashboard could address the requirements of both categories of end users. In this chapter, we have decided to give focus to domain stakeholders, as they are the main target users in the context of the APs Framework.

We will now explore existing tools that are commonly used by domain stakeholders to help in the decision-making process.

5.2 Related Work: Visualization Tools for Decision Support

In this section, we review existing visualization tools that are used for decision support in various domains. We start with the most accessible ones, such as Domain-Specific Dashboards (Subsection 5.2.1) and Business Intelligence (BI) platforms (Subsection 5.2.2), and then move on to more specialized ones like Geographic Information Systems (GIS) (Subsection 5.2.3), Linguistic Information Visualizations (Subsection 5.2.4), and Generic Visualization Libraries (Subsection 5.2.5). At the end of this section (Subsection 5.2.6), we will compare them with our requirements from Figure 5.2.

5.2.1 Domain-Specific Dashboards

One of the primary tools used in decision-making processes, sometimes based on social media analyses, is Domain-Specific Dashboards. Table 5.1 showcases examples of Domain-Specific Dashboards in various domains. These dashboards, often multidimensional, concentrate on crucial aspects known as analytical dimensions pertinent to their respective domains (see Table 5.1, *Analytical Dimensions*). The visual representations provided by these dashboards come in two main types: dynamic and interactive, or static, such as through documents or images (see Table 5.1, *Interactions*).

For instance, in the tourism domain, dashboards focus on multiple dimensions such as visitor flows (UNWTO, 2023), offerings (Isère Attractivité, 2023), employment (OECD, 2023), visitors' profiles (Pilat Tourisme, 2022) and revenue generated by tourism (Atout France, 2023). Similarly, in the healthcare industry, dashboards are developed for medical professionals (Stadler et al., 2016) or to monitor the spread of pandemics (Dong et al., 2020). The economic domain also greatly depends on dashboards for valuable indicators of purchasing power, business activities demographics (INSEE, 2023b,a), etc., helping in public policy decision-making. Numerous other domains also leverage dashboards, including food (Fanzo et al., 2020), cybersecurity (McKenna et al., 2016), urban planning (Han et al., 2020), education (Park and Jo, 2015), or finance (Flood et al., 2016).

These dashboards employ a various array of visuals, including maps, charts, tables, and quantitative statistics (see Table 5.1, *Visuals*) and multilevel navigation methods such as pages, tabs, and visual groupings (see Table 5.1, *Navigation*). They are designed to be user-friendly for individuals without extensive computing expertise. While some dashboards offer interactivity through web-based platforms, enabling users to filter data and adjust the granularity of analysis, others remain more static.

However, the primary limitation of these dashboards is their domain-specific nature, which makes them highly specialized and difficult to reuse for applications across varying domains with distinct requirements. As illustrated in Table 5.1 (*Analytical Dimensions*), the analytical dimensions in these dashboards tend to be narrowly focused and they lack genericity. To alleviate this, alternative, more generic solutions have been proposed: Business Intelligence (BI) platforms.

Reference	Domain	Analytical Dimensions	Navigation	Visuals	Interactions
Reinhart and Wienold (2011)	Architecture	*Simulation of Daylit Spaces, Daylight Availability, Comfort, Energy*	*Tabs* → *Groupings*	Tables, Various Charts, Quantitative Statistics	No (static)
Stadler et al. (2016)	Healthcare	*Sepsis Outcomes, 30-day readmissions in Hospital*	*Pages*	Tables, Bar Charts, Line Charts	No (static)
Dong et al. (2020)	Covid-19	*Number of Covid-19 cases worldwide, Evolution of Covid-19 cases*	*Groupings* → *Tabs*	Bubble Map, Tables, Quantitative Stats, Bar Charts	Yes (filtering, *granularity*)
Pilat Tourisme (2022)	Tourism in Pilat	*General attendance, Touristic Offers, Types of Accommodations, Outdoor activity facilities, Visitors' Profiles*	*Pages* → *Groupings*	Choropleth Maps, Bar Charts, Quantitative Stats, Pie Chart, Treemap	No (static)
Visit Paris Region (2023)	Tourism in Paris	*Hotel Attendance, Touristic Frequentation, Aerial Reservations, Touristic Activities*	*Pages* → *Groupings*	Line Charts, Bar Charts, Quantitative Stats, Variance Charts, Bubble Maps	Yes (filtering)
INSEE (2023a)	Local Public Policies	*Incomes and Purchasing Power, Job Market, Field of Activity, Demography, Economics*	*Pages* → *Groupings*	Bubble Maps, Choropleth Map, Quantitative Stats, Tables	Yes (filtering)
UNWTO (2023)	Global Tourism	*Visitors' Flows, Incomes and Expenditures of Tourism, Seasonality, Accommodations, Internal Tourism*	*Pages* → *Tabs* → *Groupings*	Choropleth Maps, Line Chart, Bar Charts, Quantitative Statistics	Yes (filtering)
INSEE (2023b)	French Economy	*Economics, Demography, Job Market, Companies, Purchasing Power, etc.*	*Pages* → *Tabs*	Bart Charts, Pie Charts, Tables, Choropleth Maps	Yes (filtering, granularity)
OECD (2023)	Global Tourism	*Tourism Jobs, Tourism Revenues*	*Pages*	Table	Yes (sorting)
Atout France (2023)	Tourism in France	*Revenues of Tourism and Accommodations, Aerial Flows, Sports Events*	*Pages* → *Groupings*	Table, Bar Charts, Line Charts, Quantitative Statistics	No (static)
Isère Attractivité (2023)	Tourism in Isere	*Touristic Offers, Visitors' Profiles*	*Pages* → *Groupings*	Tables, Bar Charts, Pie Charts, Quantitative Statistics	No (static)
Park and Jo (2015)	Education in Korea	*Students' Activity on a Online Platform*	*Pages*	Bar Charts, Line Charts, Points Clouds	No (static)
Biehl et al. (2007)	Team Management	*Actions on a Shared Codebase*	*Pages* → *Groupings*	Lists, Cards, Icons	No (static)
Williams et al. (2017)	Chemistry	*Chemicals Properties, Toxicity Values, Exposure, Hazard*	*Pages* → *Tabs* → *Groupings*	Quantitative Stats, Schemes, Bar Charts, Density Charts, Tables, Points Clouds, Raw Text	Yes (various actions)
McArdle and Kitchin (2016)	Urban Planning	*Housing, Social, Transport, Health, Education, Crime, Industry, Demographics*	*Pages* → *Tabs* → *Groupings*	Quantitative Stats, Tables, Various Maps, Line Charts, Stacked Charts, Pictures, Videos	Yes (filtering)
McKenna et al. (2016)	Cyber-Security	*Network Records and Alerts*	*Pages* → *Groupings*	Quantitative Stats, Heatmaps, Tables, Bubble Maps, Bar Charts	Yes (filtering)
Fanzo et al. (2020)	Food	*Food Supply Chains, Food Environments, Food Affordability, Food Security, Food Policies*	*Pages* → *Tabs* → *Groupings*	Quantitative Stats, Choropleth Maps, Tables, Bar Charts, Line Charts	Yes (filtering, granularity)

Table 5.1: Comparison of Existing Dashboards in Various Domains of Application

5.2.2 Business Intelligence (BI)

Business Intelligence (BI) has been used as an umbrella term to describe concepts and methods to improve business decision-making by using fact-based support systems (Lim et al., 2013). BI's core objective is to enhance the quality of decisions through the systematic analysis of data.

Among BI tools available, *Tableau Public* (Datig and Whiting, 2018), *Power BI* (Ferrari and Russo, 2016), *SAP BusinessObjects* (Howson et al., 2012), and *QlikView* (Shukla and Dhir, 2016) stand out for their contributions to decision-making efficacy. These platforms are celebrated for their interactive data visualization capabilities and intuitive user-built dashboards. Indeed, the latter are often created by end users themselves from an extensible library of visuals.

These dashboards facilitate data-driven decisions (Hansoti, 2010; Orlovskyi and Kopp, 2020) through their interactive data exploration and user-friendly dashboards. BI tools are characterized by their data aggregation (often combined with OLAP, *Online Analytical Processing* platforms, (Negash, 2004)), analysis, and visualization capabilities. These tools excel at processing large volumes of structured data from various sources, enabling businesses to gain actionable insights with speed and accuracy. The core features of leading BI tools include their genericity, their ability to be adapted to various domains of application, and their advanced interactivity with combined filtering, dynamic change of analysis granularity, and building of customizable dashboards. They have been used in various domains like hospitality and tourism (Mariani et al., 2018; Höpken and Fuchs, 2022), healthcare (Ashrafi et al., 2014; Cunha et al., 2023), banking (Moro et al., 2015), or manufacturing (Intelligence, 2009). Applications to social media also exist (Cuzzocrea et al., 2016).

However, BI tools are primarily designed to handle numerical and well-structured data, resulting in significant challenges when working with unstructured text data such as social media posts. While certain efforts have been made to incorporate NLP processes into BI tools (Vashisht and Dharia, 2020; Desai et al., 2021), they still struggle to present sequential and association data and draw connections across text annotations from various dimensions. Additionally, these tools lack comprehensive support for multilingual data and the visuals they offer are sometimes too generic and cannot present optimally the combined dimensions of social media data (e.g., sentiment, engagement, etc.).

BI tools can visualize spatial data, but more specialized and advanced tools exist for this purpose, such as Geographic Information Systems (GIS).

5.2.3 Geographic Information Systems (GIS)

Geographic Information Systems (GIS) are tools frequently employed by stakeholders across various domains to analyze, visualize, and interpret spatial and geographic data.

These data are often linked with other dimensions, such as thematic (e.g., crop type (Pan and Pan, 2012), industrial activities (Agbalagba et al., 2016)), or temporal aspects (e.g., the periodic evolution of these representations (Siabato et al., 2018)). GIS is capable of capturing, storing, querying, analyzing, and visualizing geographically referenced information (Chang, 2008). Geospatial data encompasses both the place (e.g., latitude and longitude, geometry) and the attributes of spatial features (Chang, 2016). These tools provide a rich suite of functionalities tailored to address the various spatial requirements of geographers, ranging from advanced, granular spatial analysis to simple mapping solutions. Table 5.2 presents an excerpt of the most commonly used GIS in decision-making processes.

Name	Type	Platform	Primary Use Case
ArcGIS	Commercial	Windows, macOS, Web-based	*Comprehensive GIS, Mapping*
QGIS	Open Source	Windows, macOS, Linux	*Mapping, Spatial Analysis*
GRASS GIS	Open Source	Windows, macOS, Linux	*Spatial Modeling, Data Processing*
Google Earth	Commercial	Web-based	*Satellite Imagery Analysis, Mapping*
MapInfo	Commercial	Windows	*Business Mapping, Data Analysis*
Global Mapper	Commercial	Windows	*Data Analysis, Management*
ERDAS IMAGINE	Commercial	Windows	*Remote Sensing, Image Processing*
gvSIG	Open Source	Windows, macOS, Linux	*GIS, Data Management*
PostGIS	Open Source	Windows, macOS, Linux	*Spatial Database Extension*
SAGA GIS	Open Source	Windows, macOS, Linux	*Spatial Analysis, Geoprocessing*

Table 5.2: Comparison of Geographic Information Systems

For domain stakeholders engaged in geography, environmental studies, urban planning, resource management, and various other spatially centered domains, GIS have proven useful in making informed decisions based on the spatial dimensions of data (Jankowski, 2009; Tahri et al., 2015). Nowadays, many GIS are available on the market, including tools designed for both power users and those seeking ease of use:

- For *geographers*: software like *QGIS* (Kurt Menke et al., 2016), *Esri's ArcGIS* (Booth et al., 2001) or *OSGeo's GRASS GIS* (Neteler and Mitasova, 2002) offer extensive features for spatial analysis, data modeling, and complex map creation. These platforms are highly valued for their robustness and flexibility in handling sophisticated geographic data layers and analyses. However, they may be difficult to use for non-geographer users.

- For *more general users*: platforms such as *Google Earth Engine* (Mutanga and Kumar, 2019) simplify the process of engaging with geospatial data, making it accessible to users without specialized training. These tools focus on providing an intuitive interface for exploring and visualizing geographic information on a global scale.

Although GIS are useful for in-depth visualization of geospatial data, as shown in Table 5.2, their focus predominantly lies in spatial data visualization, which somewhat constrains their utility for non-spatial requirements. Despite this limitation, GIS can present other types of data (e.g., thematic) (Murthy et al., 2003), provided they are linked to spatial features.

5.2.4 Linguistic Information Visualizations

The next category of tools we review is more specialized: Linguistic Information Visualizations (Penn and Carpendale, 2009). These tools can be divided into three main categories based on their accessibility (refer to Table 5.3, *Target Users*).

Some tools (see Table 5.3, tools for computer scientists), such as *SpaCy* (Chantrapornchai and Tunsakul, 2021), *TextRazor* (Rajaonarivo et al., 2022), *GATE* (Maynard et al., 2000), and *Gensim* (Rehurek and Sojka, 2011), are more oriented toward computer scientists looking to develop NLP applications. These tools primarily focus on data processing and offer limited visualization capabilities, such as text-based annotation views, semantic or dependency trees, etc (see Table 5.3, *Visualization Capabilities*). These tools are challenging for non-computer scientists to use due to their complexity and mostly require programming knowledge. They also lack interactivity and mostly produce static visualizations.

Name	Primary Focus	Visualization Capabilities	Target Users
SpaCy	Text Processing	Dependency Trees, Annotations Highlighting	Computer Scientists
TextRazor	Text Processing	Tables, Dependency Trees, Annotations Highlighting	Computer Scientists
Gensim	Text Processing	Topic Modeling and Word Embedding Graphs	Computer Scientists
GATE	Text Processing	Syntactic Graphs, Annotations Highlighting	Computer Scientists
TermoStat	Statistical Text Analysis	Tables, Annotations Highlighting, Word Clouds, Quantitative Statistics	Linguists
KonText	Statistical Text Analysis	Tables, Annotations Highlighting, Quantitative Statistics	Linguists
IRaMuTeQ	Statistical Text Analysis	Co-occurrence Graphs, Word Clouds, Annotations Highlighting, Quantitative Statistics	Linguists
Voyant Tools	Statistical Text Analysis	Tables, Word Clouds, Charts, Annotations Highlighting, Quantitative Statistics	Linguists
VOSviewer	Bibliographical Analysis	Bibliographical Graphs, Quantitative Statistics	General Users
SentimentViz	Sentiment Analysis	Graphs, Heatmaps, Maps, Scatter Plots, Tables, Bar Charts, Word Clouds	General Users

Table 5.3: Comparison of Linguistic Processing and Visualization Tools

Others (see Table 5.3, tools for computer linguists), such as *IRaMuTeQ* (Loubère and Ratinaud, 2014), *Voyant Tools* (Welsh, 2014), *TermoStat* (Terryn et al., 2019), and *KonText* (Machálek, 2020), provide a broader range of visualization options because they are targeted at non-computer scientists users, like linguists looking for statistical text analysis. They focus on word-based statistical analyses and provide visuals adapted to that, like word clouds, quantitative statistics, co-occurrence graphs, etc (see Table 5.3, *Visualization Capabilities*). They require deep knowledge of languages, with complex analyses, and therefore are not fit for non-linguist users.

Finally (see Table 5.3, tools for general users), some tools are very user-friendly with various visualizations modules like *VOSviewer* (Van Eck and Waltman, 2013) for bibliographical analyses, and *SentimentViz* (Healey and Ramaswamy, 2022) for sentiment analysis. These tools are easy to use, and interactive but generally focus on a single dimension (such as *sentiment* or *a specific domain of interest*) and therefore limited for (see Table 5.3, *Primary Focus*).

5.2.5 Generic Visualization Libraries

Lastly, we acknowledge the existence of various generic visualization libraries that help build static, animated, and interactive dashboards for a broad array of data type.

Libraries such as *Matplotlib* (Tosi, 2009), *Chart.js* (Da Rocha, 2019), and *Plotly* (Sievert, 2020) are instrumental in this regard. Furthermore, for the visualization of geospatial data, *Leaflet* (Crickard III, 2014) and *Folium* (Gupta and Bagchi, 2024) enable the representation of data overlays on maps, enhancing the spatial analysis capabilities. In the domain of 3D visualizations, *Three.JS*

(Kenwright, 2019) offers a powerful foundation for designing and displaying animated 3D computer graphics directly in the web browser. When it comes to visualizing interconnected data and relationships, *Cytoscape* (Smoot et al., 2011) and *Gephi* (Bastian et al., 2009) specialize in graph and network visualization, facilitating the understanding of complex networks through intuitive visual representations. Lastly, *D3.JS* (Zhu, 2013) distinguishes itself through its flexibility and control over the final visual outcome, enabling the creation of highly complex and responsive visualizations.

However, it is important to note that these libraries are either highly technical or require programming knowledge, making them less accessible to domain stakeholders who are not computer scientists. Despite this, they remain valuable tools for designing domain-adaptable dashboards, offering a rich set of resources for visualizing various data types.

We will now evaluate how these families of tools address the requirements we have defined within the context of our framework (see Figure 5.2).

5.2.6 Discussion and Comparison with our Requirements

Figure 5.3 presents a visual summary of the strengths and limitations of the tools reviewed in this section. As we can see, most of the tools' families present interesting features, but they also have severe limitations preventing their direct applicability in the context of our framework.

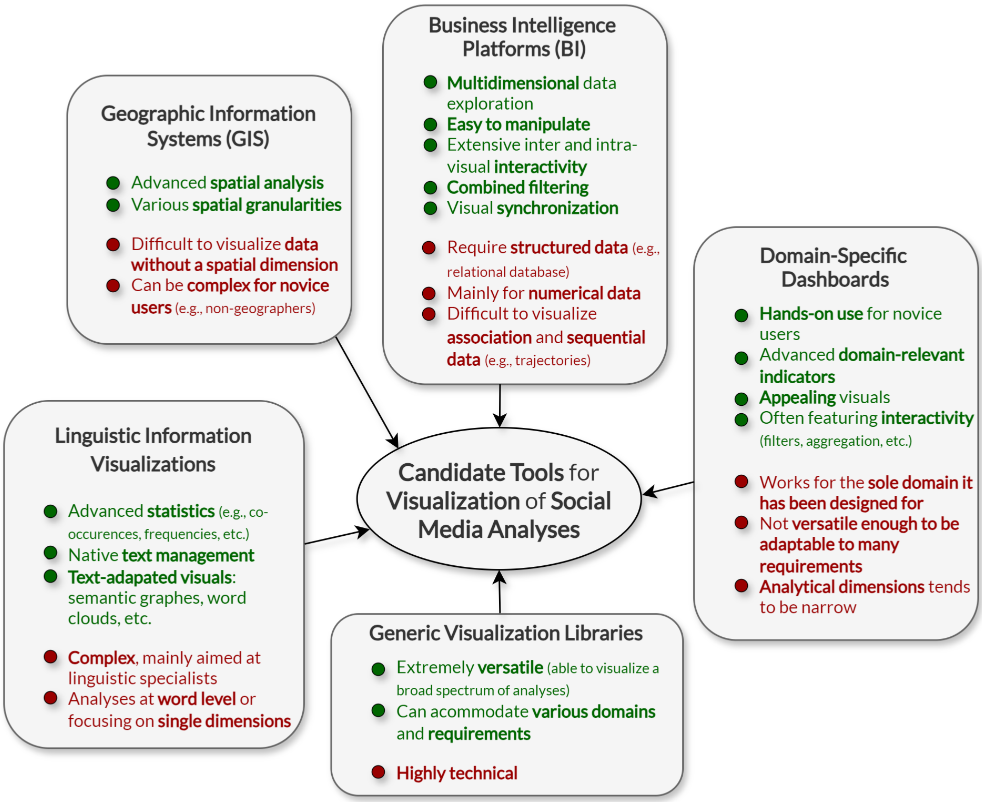

Business Intelligence Platforms (BI)
● Multidimensional data exploration
● Easy to manipulate
● Extensive inter and intra-visual interactivity
● Combined filtering
● Visual synchronization
● Require structured data (e.g., relational database)
● Mainly for numerical data
● Difficult to visualize association and sequential data (e.g., trajectories)
Geographic Information Systems (GIS)
● Advanced spatial analysis
● Various spatial granularities
● Difficult to visualize data without a spatial dimension
● Can be complex for novice users (e.g., non-geographers)
Domain-Specific Dashboards
● Hands-on use for novice users
● Advanced domain-relevant indicators
● Appealing visuals
● Often featuring interactivity (filters, aggregation, etc.)
● Works for the sole domain it has been designed for
● Not versatile enough to be adaptable to many requirements
● Analytical dimensions tends to be narrow
Linguistic Information Visualizations
● Advanced statistics (e.g., co-occurences, frequencies, etc.)
● Native text management
● Text-adapated visuals: semantic graphes, word clouds, etc.
● Complex, mainly aimed at linguistic specialists
● Analyses at word level or focusing on single dimensions
Candidate Tools for Visualization of Social Media Analyses
Generic Visualization Libraries
● Extremely versatile (able to visualize a broad spectrum of analyses)
● Can acommodate various domains and requirements
● Highly technical

Figure 5.3: Summary of the Strengths and Limitations of the Candidate Visualization Tools Reviewed

Table 5.4 shows a detailed view of the requirements from Figure 5.2 displayed in a tabular way and correlated with the families of visualization tools we have just reviewed.

We have used a tricolor heatmap. A red color indicates that the tool family is not fit to address the requirement at all, an orange color signifies that the tool family can address the requirement but with major shortcomings and limitations, while a green color indicates that the tool family tends to be suitable for the requirement.

		Domain-Specific Dashboards	Business Intelligence Platforms	Geographic Information Systems	Linguistic Information Visualizations	Generic Visualization Libraries
Genericity						
Domain-Adaptable		🟥	🟩	🟩	🟧	🟩
Source-Independent		🟥	🟧	🟧	🟩	🟩
Versatile		🟧	🟧	🟧	🟧	🟩
Domain Stakeholders	**NLP Researchers**					
Observe						
Satisfaction	*Sentiment*	🟩	🟩	🟧	🟩	🟧
Places	*Location Entities*	🟩	🟩	🟩	🟩	🟧
Topics	*Thematic Entities*	🟩	🟩	🟧	🟩	🟧
Time	*Temporal Entities*	🟩	🟩	🟧	🟧	🟧
Engagement		🟩	🟩	🟧	🟥	🟧
Users and Profiles		🟩	🟩	🟧	🟥	🟧
	Text Metadata	🟩	🟩	🟥	🟩	🟧
	Raw Text	🟩	🟩	🟥	🟩	🟧
	Text Annotations	🟩	🟥	🟥	🟩	🟧
	Token Annotations	🟩	🟥	🟥	🟩	🟧
Indicators						
Frequencies	*Annotation Distribution*	🟩	🟩	🟧	🟩	🟧
Associations	*Annotation Co-occurrence*	🟩	🟥	🟧	🟩	🟧
Trajectories	*Annotations Sequencings*	🟩	🟥	🟩	🟩	🟧
Various Statistics	*Annotations Statistics*	🟩	🟩	🟧	🟩	🟧
Multi-Criteria Indicators		🟧	🟧	🟧	🟥	🟧
Interactivity						
Filter		🟩	🟩	🟩	🟩	🟧
Aggregate		🟩	🟩	🟩	🟧	🟧
Change Granularity		🟩	🟩	🟩	🟥	🟧
Adapt Indicators		🟧	🟧	🟧	🟥	🟧
Easy to Use		🟩	🟩	🟧	🟧	🟥
	Extensibility	🟥	🟩	🟩	🟧	🟩

Table 5.4: Comparison of Reviewed Visualization Tools with our Visualization Requirements

To design our generic dashboard, we will be extending, integrating, and blending features from a variety of these tools. The findings of Table 5.4 are as follows:

- *Domain-Specific Dashboards* can potentially fulfill nearly all our requirements but only within the specific domain for which they were developed, which contradicts the philosophy of domain-independence of our framework. They are also rarely extensible, which means it is not easy to incorporate new analytical dimensions into them easily.
- *Business Intelligence Platforms* could provide a solid basis for our generic dashboard as they are adaptable across domains, interactive (various filtering and data aggregation functions), user-friendly (even for non-computer scientist users), and capable of visualizing and calculating many of the required indicators. However, they face challenges in presenting sequence and association data and rely on natively numerical rather than text-based indicators, which limits their applicability to our requirements. From BI tools, we will borrow the interactive design, user-friendly interfaces, and visual synchronization, adapting these features for their use with annotated social media posts as opposed to traditional numerical data.
- *Geographic Information Systems* are constrained by their spatial focus. They can consider other dimensions but only when linked with a spatial one (for example, the weather in different regions). We will adopt the detailed multi-granularity spatial views of GIS, acknowledging that social media data often contains a spatial aspect. However, unlike traditional GIS, we aim for a multidimensional approach that extends beyond merely spatial focus, and we do not require as comprehensive spatial analyses.
- *Linguistic Information Visualizations* excel at computing text-based indicators but may be challenging for non-linguist users to manipulate, struggle with non-text based data (like social media posts' metadata), and have limited interactivity and no ability to build custom indicators. From *Linguistic Information Visualizations*, we will take their analytical strength, such as co-occurrence and frequency analysis, but go beyond their usual focus on text and words to include dimensional entities extracted from posts' content and metadata like thematic concepts, places, periods, sentiment, engagement, etc.
- *Generic Visualization Libraries* could potentially meet all our requirements but require programming knowledge and are therefore not suitable for non-computer scientists' end users. We will, however, build a generic dashboard leveraging them.

By combining these elements, our generic dashboard aims to provide an inclusive visualization of social media analyses that benefits both NLP researchers and domain stakeholders (refer to Figure 5.2). The objective is to achieve a result close to domain-specific dashboards but while being domain-independent, source-independent, and highly versatile. We will now present how we extended, integrated, and blended features of these existing visualization tools to build *TextBI* (Masson et al., 2024c), a generic, domain-adaptable dashboard to visualize social media analyses.

5.3 *TextBI* Dashboard Structure and Conceptual Aspects

TextBI (overview of the structure in Figure 5.4) is a generic dashboard that allows for interactive visualization of social media analyses (annotations and indicators). It is designed to be adaptable to any social media source and domain of application, provided the data adheres to the generic APs Trajectory Model (see Chapter 4).

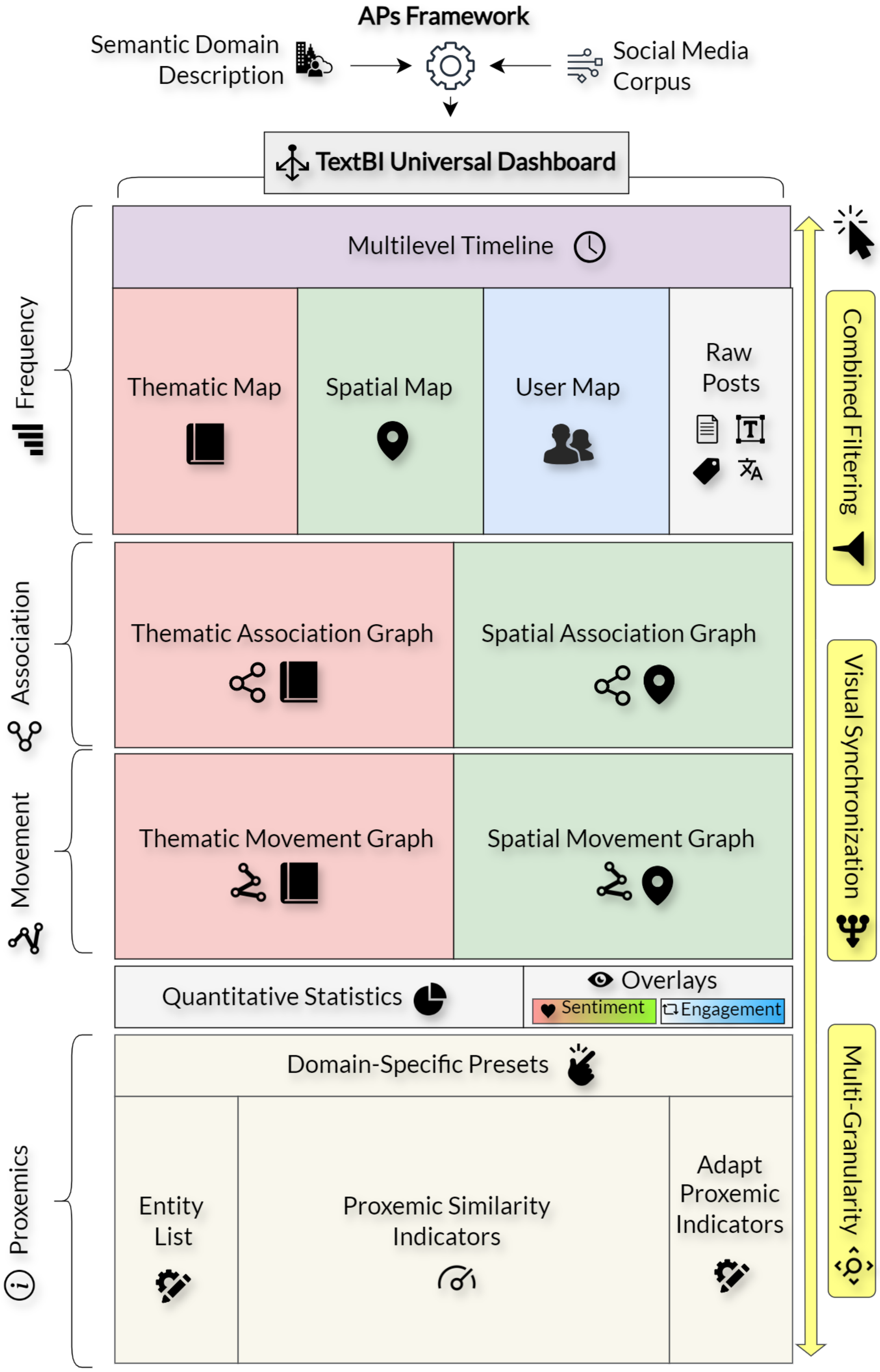


Figure 5.4: Architecture of the *TextBI* Dashboard. Icons are mapped to requirements from Figure 5.2

The dashboard offers a variety of features to facilitate the analysis of multilingual social media data across several core dimensions.

In Figure 5.4, we reuse the **same set of icons** as in the requirements (presented in Figure 5.2) to highlight which parts of the dashboard address specific requirements. The dashboard takes as input the model resulting from the previous phases of the framework, namely, any semantic domain description and any corpus of social media posts processed using the APs Framework (*Collect*, *Transform*, and *Analyze*). Refer to Figure 5.4 for the *Semantic Domain Description*, *APs Framework*, and *Social Media Corpus* items.

The dashboard is structured around four main views: *Frequency* (see Subsection 5.3.1 and Figure 5.4, *Frequency*), *Association* (see Subsection 5.3.2 and Figure 5.4, *Association*), *Movement* (see Subsection 5.3.3 and Figure 5.4, *Movement*), and *Proxemics* (see Subsection 5.3.4 and Figure 5.4, *Proxemics*), corresponding to the indicators required from each category of users (presented in Figure 5.2, *Indicators*). Additionally, a *Statistics* view and superimposed *Overlays* are available (see Subsection 5.3.5). For now, we will present the theoretical aspects of the dashboard and we will apply it to the domain of tourism in the next section (Section 5.4).

5.3.1 Frequency View

The *Frequency* view (Figure 5.4, *Frequency* and Figure 5.5) serves as the primary interface, highlighting spatial, temporal, and thematic frequencies through four major visualizations. It caters to the requirements of both categories of users from Figure 5.2, namely, by visualizing frequencies and distributions of multidimensional annotations.

Note that in Figure 5.5 we feature the type of visual used for each dimension (e.g., treemap, scatter plot, etc.) but also the main inspiration we have used when designing the visual (refer to Section 5.2, DSD corresponds to Domain-Specific Dashboards, BI to Business Intelligence, GIS to Geographic Information Systems, and LV to Linguistic Information Visualizations).

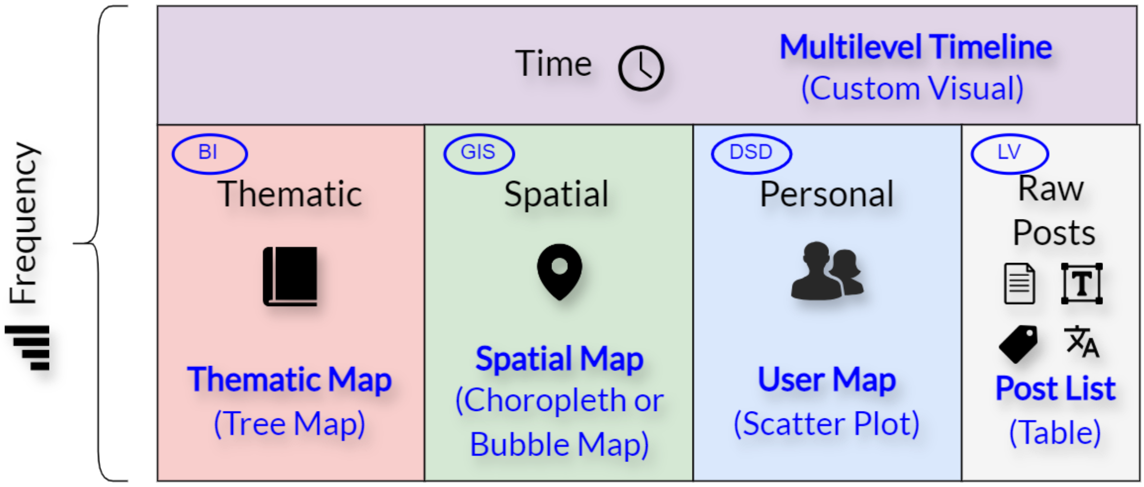


Figure 5.5: Overview of the Frequency View

- The *Thematic Map* (see Figure 5.5, *Thematic*) is inspired from treemaps featured in BI tools. It visualizes the hierarchical structure and frequency of thematic concepts from the semantic resource (such as a dictionary, thesaurus, or ontology) applied to the social media corpus.

The granularity of this visual can be adjusted; for instance, it may be set to consider thematic concepts only at specific levels within a thesaurus or ontology.

- The *Spatial Map* (see Figure 5.5, *Spatial*) has two modes inspired by GIS (e.g., choropleth and proportional symbol maps (Słomska-Przech and Gołębiowska, 2021), which in our case is a bubble map). It showcases the frequency of places mentioned in posts. Users can set the spatial granularity, which can range from broad categories such as countries to more specific ones like Points of Interest (POIs). Places are aggregated depending on the chosen granularity. The map uses a linear gradient or bubble size for representation, with more transparent areas signifying fewer originating posts. In the same way as the *Thematic Map* the granularity of analysis can be changed to fit the domain and type of data used.

- The *User Map* (see Figure 5.5, *Personal*) is represented as a scatter plot. This kind of chart is often featured in domain-specific dashboards to correlate various variables. In our case, it presents the users' posting frequency on the x-axis against the count of the users' followers on the y-axis. This design helps in the rapid identification of influential users. Each user's language is represented by color and symbol, with the symbol's size corresponding to the number of posts from that user. We are looking to make the colors and axis customizable by the end users.

- The *Timeline* (see Figure 5.5, *Time*) visual offers a visualization of the volume of posts per day across the dataset range, segmented into different times of the day such as morning, afternoon, or evening. It provides various temporal granularity options, including days, months, seasons, and years. This is a custom visual we have designed from scratch that did not exist in the existing tools we have reviewed.

- The *Post List* (see Figure 5.5, *Raw Posts*) allows to visualize raw social media posts without any processing. It is used to contextualize the other visualizations and is inspired by Linguistic Information Visualization. This was a recurrent requirement of NLP researchers (Figure 5.2, *Raw Text*, *Text and Token Annotations*).

5.3.2 Association View

The *Association* view (Figure 5.4, *Association* and Figure 5.6) presents visuals that illustrate the connections between entities (e.g., places and themes) through their co-occurrences in posts.

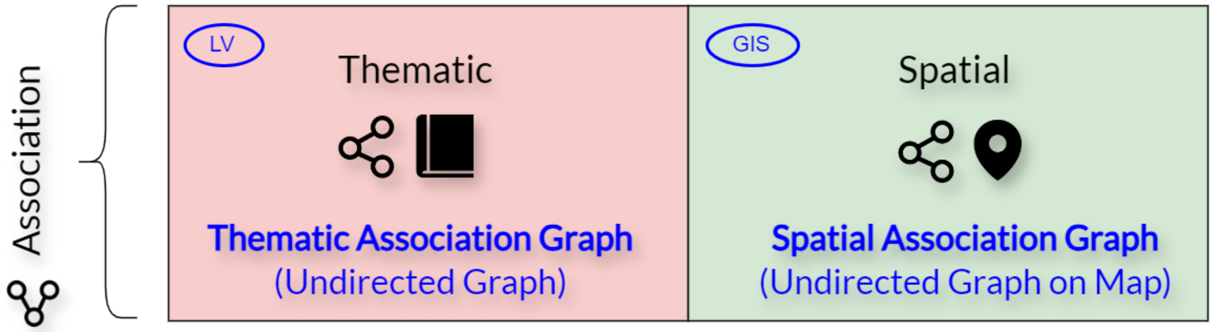

Association
LV
Thematic
Thematic Association Graph
(Undirected Graph)
GIS
Spatial
Spatial Association Graph
(Undirected Graph on Map)

Figure 5.6: Overview of the Association View

This is also a requirement from both categories of users from Figure 5.2 (*Association* and *Annotations Co-occurrence*). These connections are depicted using undirected graphs, where the nodes represent entities such as thematic concepts or places, and the edges indicate the strength of co-occurrence between them. This allows for easy identification of heavily correlated concepts or places.

- The *Thematic Association Graph* (see Figure 5.6, *Thematic Association Graph*) visualizes thematic associations as an undirected graph with nodes representing thematic concepts. The size of the nodes is proportional to the frequency of occurrences in the social media dataset, and the thickness of the edges is proportional to the number of co-occurrences between the concepts (e.g., several concepts appearing in the same posts). This visual is inspired by Linguistic Information Visualization (like *Iramuteq* (Loubère and Ratinaud, 2014)) but simplified for use by non-linguist users. This visual allows observers to note associated concepts.

- The *Spatial Association Graph* (see Figure 5.6, *Spatial Association Graph*) works similarly, but this time the graph is superimposed on a spatial map. Nodes represent places and are placed accordingly on the map, allowing easy identification of places that are correlated in posts. Such visualizations are possible with GIS but require a lot of configuration beforehand and are therefore not easily accessible to domain stakeholders.

5.3.3 Movement View

Aggregated trajectories or sequences can be displayed in the *Movement* view (Figure 5.4, *Movement* and Figure 5.7), focusing on the sequencing of entities in user trajectories, for example, the transition from one thematic concept or place to another. This sequencing is visualized through directed graphs where edges indicate the amount of time two concepts or places are sequenced in user posts' trajectories (e.g., appearing in two contiguous posts).

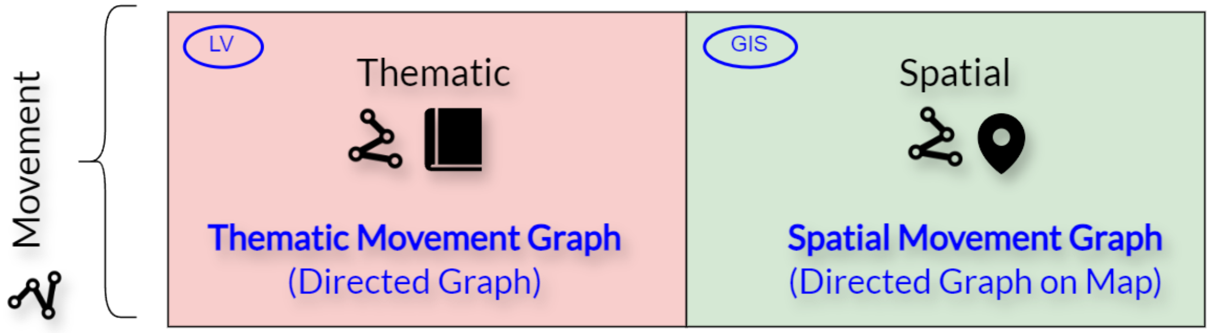


Figure 5.7: Overview of the Movement View

- The *Thematic Movement Graph* (see Figure 5.7, *Thematic Movement Graph*) displays aggregated sequences of thematic concepts. It highlights thematic concepts that are often chained between posts.

- The *Spatial Movement Graph* (see Figure 5.7, *Spatial Movement Graph*) works similarly, but this time the graph is again superimposed on a spatial map. Nodes represent places and are placed accordingly on the map, allowing easy identification of dominant spatial flows between places.

5.3.4 *Proxemics* View

The *Proxemics* view is different. It addresses the requirement for multi-criteria, adaptable indicators from domain stakeholders (presented in Figure 5.2, *Multi-Criteria Indicators* and *Adapt Indicators*) by visually presenting proxemic analyses. It is interfaced with the *ProxMetrics* toolkit introduced in Chapter 4 to calculate proxemic similarity indicators. The *Proxemics* view allows the setup of domain-adaptable indicators, their adaptation, and finally their visualization. It is a novel visual that we have not seen in any existing tools we have reviewed.

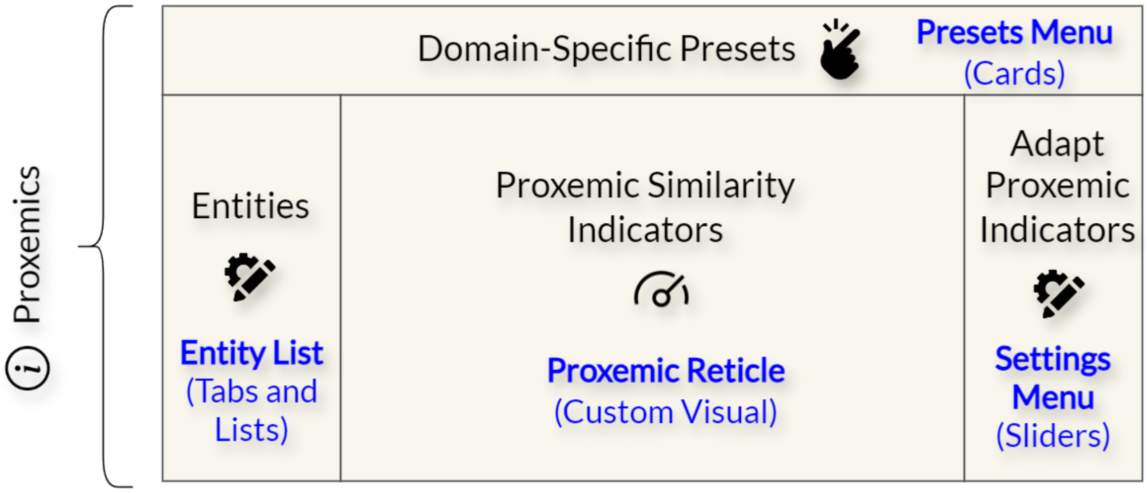


Figure 5.8: Overview of the *Proxemics* View

- The *Entity List* (see Figure 5.8, *Entity List*) allows the selection of the reference entity (as a reminder, *proxemics* requires the selection of a *reference entity*, also called the *center entity*). Various lists are distributed in tabs for each category of entities (e.g., users, groups, places, periods, themes). A search feature is provided. Each entity can be dragged and dropped at the center of the *Proxemic Reticle*.

- The *Proxemic Reticle* (see Figure 5.8, *Proxemic Reticle*) is positioned at the center. It allows the visualization of the reference entity at the center with target entities scattered around it in various proxemic zones depending on their proxemic similarity to the reference one. The type of target entities can be changed dynamically at will.

- The *Settings Menu* (see Figure 5.8, *Settings Menu*) allows tweaking of the five proxemic dimensions in use (Distance, Identity, Location, Movement, Orientation, or DILMO) through five sliders, giving more or less importance to certain dimensions in order to create personalized indicators according to the domain. Any action in this menu dynamically updates the *Proxemic Reticle* without a reload.

- The *Presets Menu* (see Figure 5.8, *Presets Menu*) consists of various presets specific to the domain. A preset is a pre-configuration of the view, for example, preconfigured sliders for each proxemic dimension, presetting of the target entities, allowing only certain reference entities (such as themes belonging only to a certain branch of a semantic resource, etc.). The objective of presets is to make the *Proxemics* view accessible to novice users without any background or training with the platform.

5.3.5 Statistics View and Overlays

TextBI's base features a *Statistics* view (Figure 5.4, *Quantitative Statistics* and Figure 5.9) showing quantitative statistics related to active filters, including post, user, concept, and place counts; current post time range; total engagement level; and prevailing sentiments.

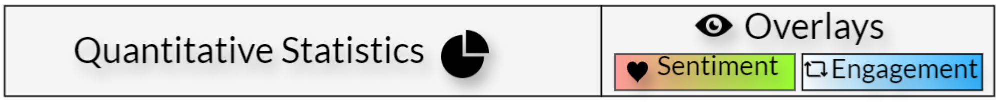


Figure 5.9: Overview of the Statistics View

Additionally, *TextBI*'s visuals support the superimposition (referred to as *Overlays*) of sentiment and engagement enrichment data (Figure 5.4, *Enrichment Overlays*). This was a requirement from both categories of users (Figure 5.2, *Satisfaction*, *Sentiment* and *Engagement*).

- The *Sentiment Overlay* is indicated through color coding (*green* for positive, *red* for negative, and *orange* for mixed sentiment), enabling a better understanding of aggregate sentiment by themes, places, periods or users.
- The *Engagement Overlay* is visualized using a linear gradient (*darker blue* indicating strong engagement, *lighter blue* indicating low engagement), providing insights into user engagement (e.g., likes, reposts, etc.).

5.3.6 Interactivity and Visual Synchronization

The *TextBI* platform employs interactions commonly found in BI (see Figure 5.4, various interactions displayed in yellow, on the right) to satisfy interactivity requirements from both categories of users (Figure 5.2, *Interactivity*), ensuring a fluid synchronization of visuals.

It accommodates multidimensional filtering options such as *spatial-temporal*, *spatial-thematic*, or *user-thematic-temporal*. When a user selects a particular place, theme, user, or time range, all subsequent visualizations adjust to display only posts associated with the chosen filter. The system even allows for combined filtering. Within its *proxemics* view, users can conveniently drag and drop a reference onto the center of the crosshair.

Additionally, *TextBI* features several shortcut interactions aimed at making interactions and filtering easier and less time-consuming for non-computer scientist users. For example, the ability to select several items simultaneously for filtering. We will elaborate on that in the experiment. All the visuals are resizable so the user can partially customize the dashboard.

5.3.7 Technical Aspects and Limitations

TextBI is a web-based dashboard developed using HTML, CSS, and *JavaScript*. It operates solely on the client side, eliminating the requirement for a back-end web server, which allows it to run locally. The data model is implemented in the JSON format. The application uses several libraries, including *Plotly* for charts, *Leaflet* for spatial maps, *Cytoscape* for graphs, *Split.js* for multi-window screens, and *Turf.js* for *GeoJSON* handling.

TextBI serves as a data display and aggregation tool, providing statistical analyses and calculating similarities. It does not engage in any data processing tasks of its own. Previous APs Framework phases like *Collect* and *Transform*, including NLP, have to be completed in advance. Currently, *TextBI* does not support any analytical dimensions beyond those mentioned above. In the future, we aim to make it easily extensible through a plugin system.

We have described the architecture of the generic dashboard. We will now experiment with it in the domain of tourism.

5.4 *TextBI* Experiment on the Tourism Domain

In the following examples, we will demonstrate each view and its visuals by loading the corpus of touristic tweets introduced in previous chapters into the *TextBI* dashboard. A demonstration video of *TextBI* is provided at the following address to facilitate readers' understanding.

`maxime-masson.github.io/TextBI` ↗ (click to watch)

As a reminder, this multilingual corpus covers the *French Basque Coast* region during the summer of 2019 and its tweets are mapped to the *Thesaurus on Tourism and Leisure Activities of the World Tourism Organization* (World Tourism Organization, 2002). The APs Trajectory Model (see Figure 4.5 and Figure 4.6) was instantiated using two types of data: metadata and tweet content. Metadata-based instantiation was a straightforward process, encompassing elements like engagement metrics, profile features, timestamps, and geotags, among others. The textual content of the tweets was processed using NLP modules to generate automatic annotations at the token level for (1) places and (2) thematic concepts, and at the text level for (3) sentiments (refer to Chapter 3). Note that an application of *TextBI* to another domain of application (e.g., *local public policies*) using a different data source (e.g., *municipality review platforms*) is presented in Chapter 6.

5.4.1 Frequency View

Here, we experiment with the *Frequency* view (see Figure 5.10) on the tourism dataset. This view displays the distribution of entities in the currently loaded dataset, namely *themes*, *places*, *dates*, and *users*. Numbers featured in the listing below correspond to those in Figure 5.10.

(1) The *Thematic Map* shows that when mapping tweet concepts to the *Thesaurus on Tourism and Leisure Activities* (World Tourism Organization, 2002), we find tourism heritage-related concepts are most frequent, accounting for roughly 40% of discovered concepts (depicted in blue), with many linked to natural resources such as the coast and sea.

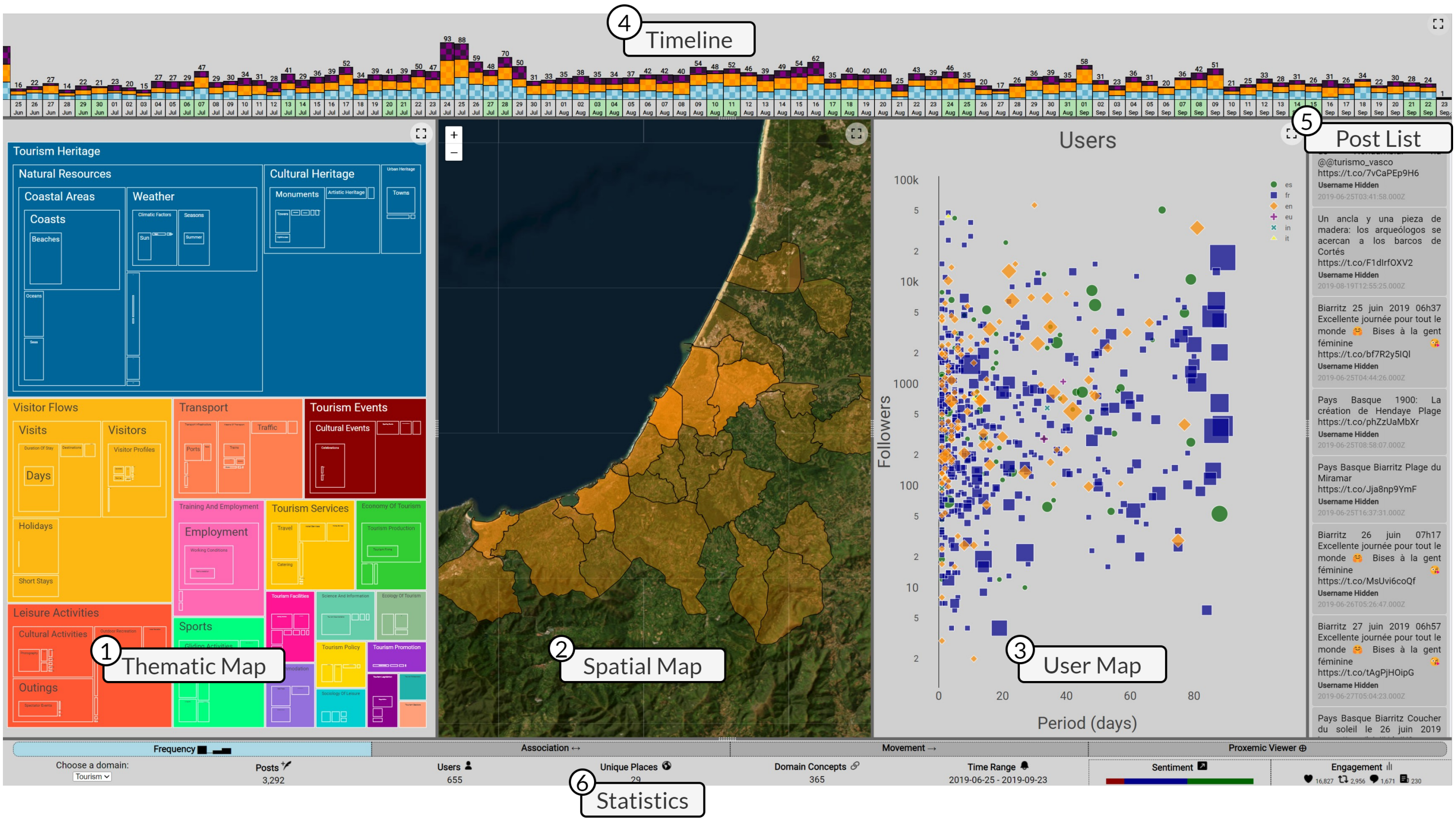
(2) The *Spatial Map* is set up at the municipality level and as a choropleth map. We observe a hotspot of tweets in three municipalities in the northernmost part of the region. These nearby municipalities appear to be popular among visitors.

(3) The *User Map* shows 655 users, spanning over 5 languages. French are depicted in *blue*, Spanish in *green*, English in *orange*, Basque in *purple*, and Italian in *yellow*.

(4) The *Timeline* offers a visualization of the volume of posts per day across the dataset range, segmented into different times of the day such as morning (depicted in *cyan*), afternoon (depicted in *orange*), and evening (depicted in *purple*). We observe a peak of tweets between the 24th and 28th of July.

(5) The *Post List* shows the currently loaded posts. Currently, no filtering is applied, so it contains all of the tweets.

(6) The *Statistics* view shows the number of tweets currently loaded, that they belong to 655 users covering 29 unique municipalities and 365 unique concepts from the tourism thesaurus in the summer of 2019. The dominant sentiment tends to be neutral to positive with a minority of negative. There were 16,827 likes, 2,956 reposts, 1,671 replies, and 230 quotes recorded in the tourist dataset.



@@turismo_vasco https://t.co/7vCaPEp9H6
Username Hidden
2019-06-25T03:41:58.000Z

Un ancla y una pieza de madera: los arqueólogos se acercan a los barcos de Cortés https://t.co/F1dIrfOXV2
Username Hidden
2019-08-19T12:55:25.000Z

Biarritz 25 juin 2019 06h37 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/bf7R2y5IQl
Username Hidden
2019-06-25T04:44:26.000Z

Pays Basque 1900: La création de Hendaye Plage https://t.co/phZzUaMbXr
Username Hidden
2019-06-25T08:58:07.000Z

Pays Basque Biarritz Plage du Miramar https://t.co/Jja8np9YmF
Username Hidden
2019-06-25T16:37:31.000Z

Biarritz 26 juin 07h17 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/MsUvi6coQf
Username Hidden
2019-06-26T05:26:47.000Z

Biarritz 27 juin 2019 06h57 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/tAgPjHOipG
Username Hidden
2019-06-27T05:04:23.000Z

Pays Basque Biarritz Coucher du soleil le 26 juin 2019

Choose a domain: Tourism
Posts: 3,292
Users: 655
Unique Places: 29
Domain Concepts: 365
Time Range: 2019-06-25 - 2019-09-23
Sentiment
Engagement: ♥ 16,827 🔁 2,956 💬 1,671 230

Figure 5.10: Screenshot of the Frequency View

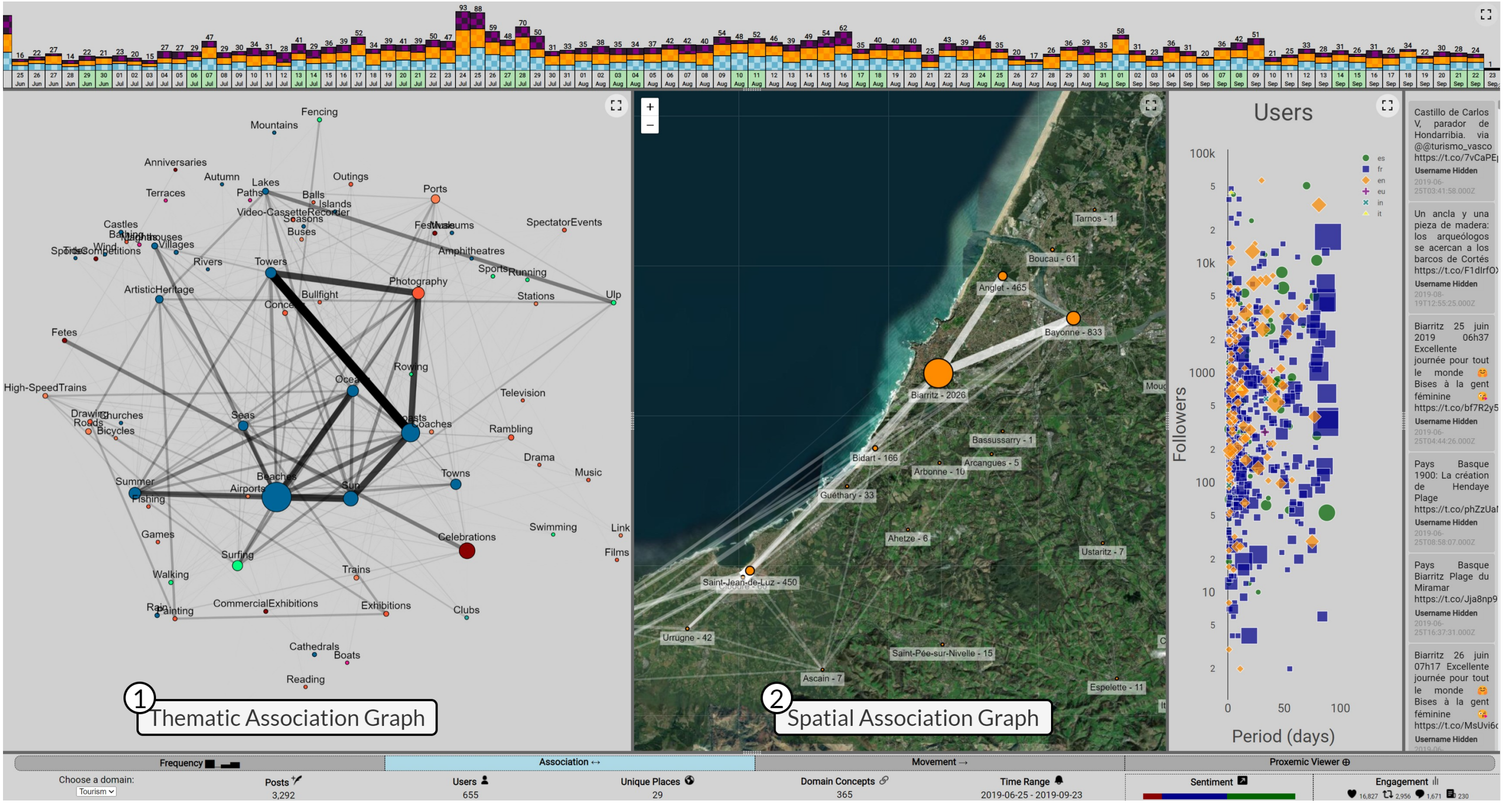
5.4.2 Association View

Figure 5.11 displays the *Association* view. This view focuses on the co-occurrence of themes and places to analyze associated entities.

(1) The *Thematic Association Graph* shows thematic concepts that are often mentioned together in

our dataset's tweets. As expected, *Sun*, *Beach*, *Surfing*, *Sea*, and other coastal concepts are heavily linked (thicker edges).

(2) The *Spatial Association Graph* shows municipalities often referenced together. These municipalities are superimposed on a spatial map for georeference. As we can see, *Biarritz* and *Bayonne* are the most associated municipalities. We hypothesize that it is because they are very close and therefore visitors often visit them on the same day and therefore reference them in the same tweets.



Castillo de Carlos V, parador de Hondarribia. via @@turismo_vasco https://t.co/7vCaPE
Username Hidden
2019-06-25T03:41:58.000Z

Un ancla y una pieza de madera: los arqueólogos se acercan a los barcos de Cortés https://t.co/F1dIrfO
Username Hidden
2019-08-19T12:55:25.000Z

Biarritz 25 juin 2019 06h37 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/bf7R2y5
Username Hidden
2019-06-25T04:44:26.000Z

Pays Basque 1900: La création de Hendaye Plage https://t.co/phZzUal
Username Hidden
2019-06-25T08:58:07.000Z

Pays Basque Biarritz Plage du Miramar https://t.co/Jja8np9
Username Hidden
2019-06-25T16:37:31.000Z

Biarritz 26 juin 07h17 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/MsUvi6
Username Hidden

Frequency | Association ↔ | Movement → | Proxemic Viewer ⊕

Choose a domain: Tourism
Posts: 3,292
Users: 655
Unique Places: 29
Domain Concepts: 365
Time Range: 2019-06-25 - 2019-09-23
Sentiment
Engagement: ♥ 16,827 2,956 1,671 230

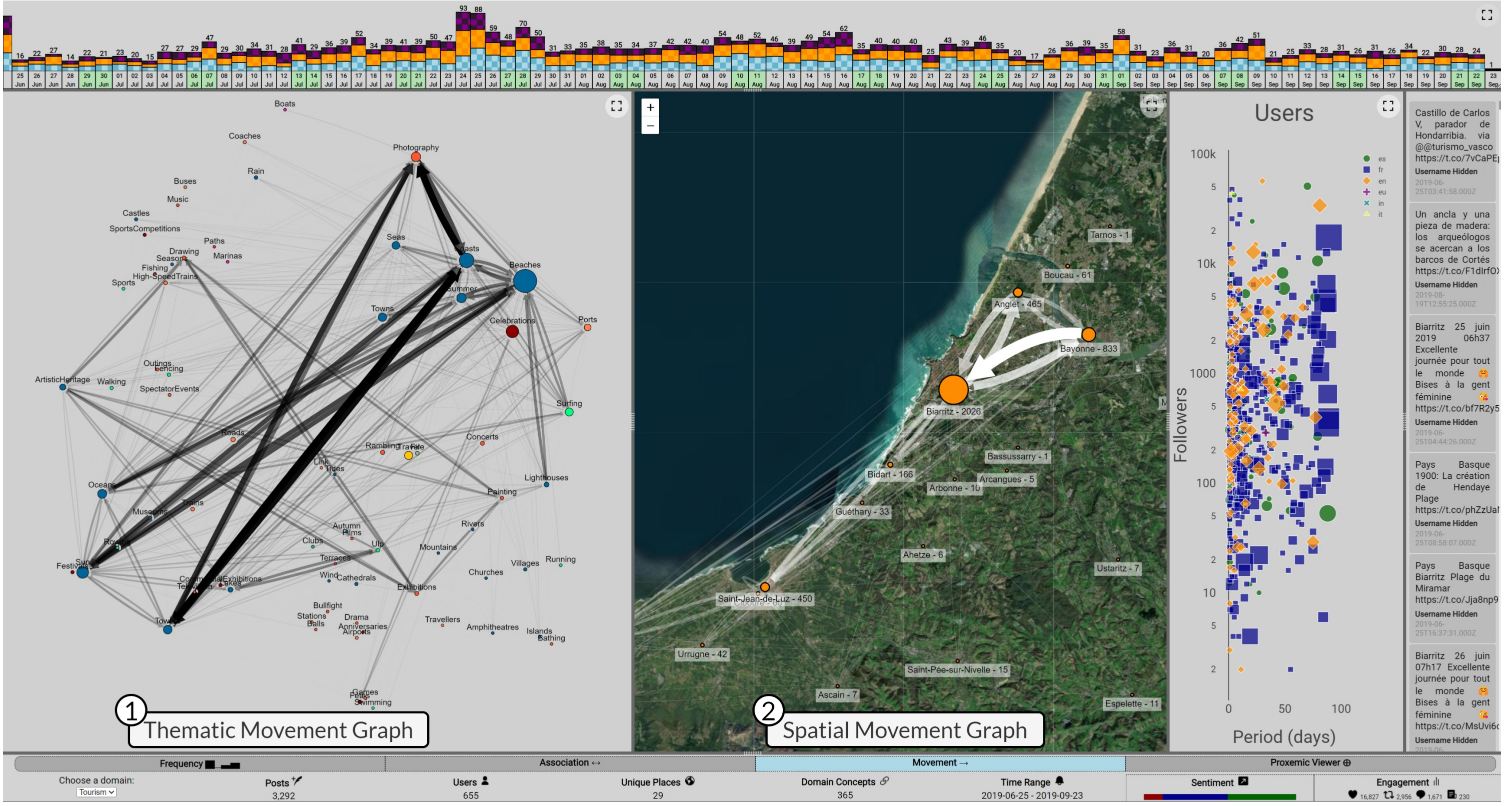
Figure 5.11: Screenshot of the Association View

5.4.3 Movement View

Figure 5.12 displays the *Movement* view. This view focuses on sequences in posts.

(1) The *Thematic Movement Graph* presents thematic sequencing. In this case, that is which touristic concepts users usually come from and go to. As we can see, there is a strong sequencing between the concepts *Coast* and *Photography*. This denotes that visitors often post about the coast, and later about taking photography.

(2) The *Spatial Movement Graph* shows visitor flows in the *French Basque Coast Area*. We can easily identify the hotspots, the main trajectories, and corridors used by visitors by looking at the size of nodes and the thickness of edges. Most visitor flows are going from *Bayonne* to *Biarritz* which are two of the most touristic municipalities in the area.



Castillo de Carlos V, parador de Hondarribia. via @@turismo_vasco https://t.co/7vCaPE
Username Hidden
2019-06-25T03:41:58.000Z

Un ancla y una pieza de madera: los arqueólogos se acercan a los barcos de Cortés https://t.co/F1dIrfOX
Username Hidden
2019-08-19T12:55:25.000Z

Biarritz 25 juin 2019 06h37 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/bf7R2y5
Username Hidden
2019-06-25T04:44:26.000Z

Pays Basque 1900: La création de Hendaye Plage https://t.co/phZzUal
Username Hidden
2019-06-25T08:58:07.000Z

Pays Basque Biarritz Plage du Miramar https://t.co/Jja8np9
Username Hidden
2019-06-25T16:37:31.000Z

Biarritz 26 juin 07h17 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/MsUvi6c
Username Hidden

Choose a domain:	Posts	Users	Unique Places	Domain Concepts	Time Range	Sentiment	Engagement
Tourism	3,292	655	29	365	2019-06-25 - 2019-09-23		♥ 16,827 🔁 2,956 💬 1,671 230

Figure 5.12: Screenshot of the Movement View

5.4.4 *Proxemics* View

The *Proxemics* view of *TextBI* (see Figure 5.13) provides users with the capability to analyze social media datasets through a proxemic lens (Hall et al., 1968; Greenberg et al., 2011). This view leverages the *ProxMetrics* toolkit and formula introduced in Chapter 4 to build custom multi-criteria indicators. It enables the visual selection of entities such as *users*, *groups*, *thematic concepts*, *places*, or *periods* by dragging and dropping them onto the main proxemic crosshair. The interface is organized with:

(1) An *Entity List* panel for the selection of the reference (center) entity.

(2) A central *Proxemic Reticle* displaying results (e.g., the proxemic similarity of target entities compared to the reference one). The type of target entities can be set up at the top left of the panel (here, depicted in green). This visual supports various combinations, including *user-to-themes*, *place-to-user*, etc.

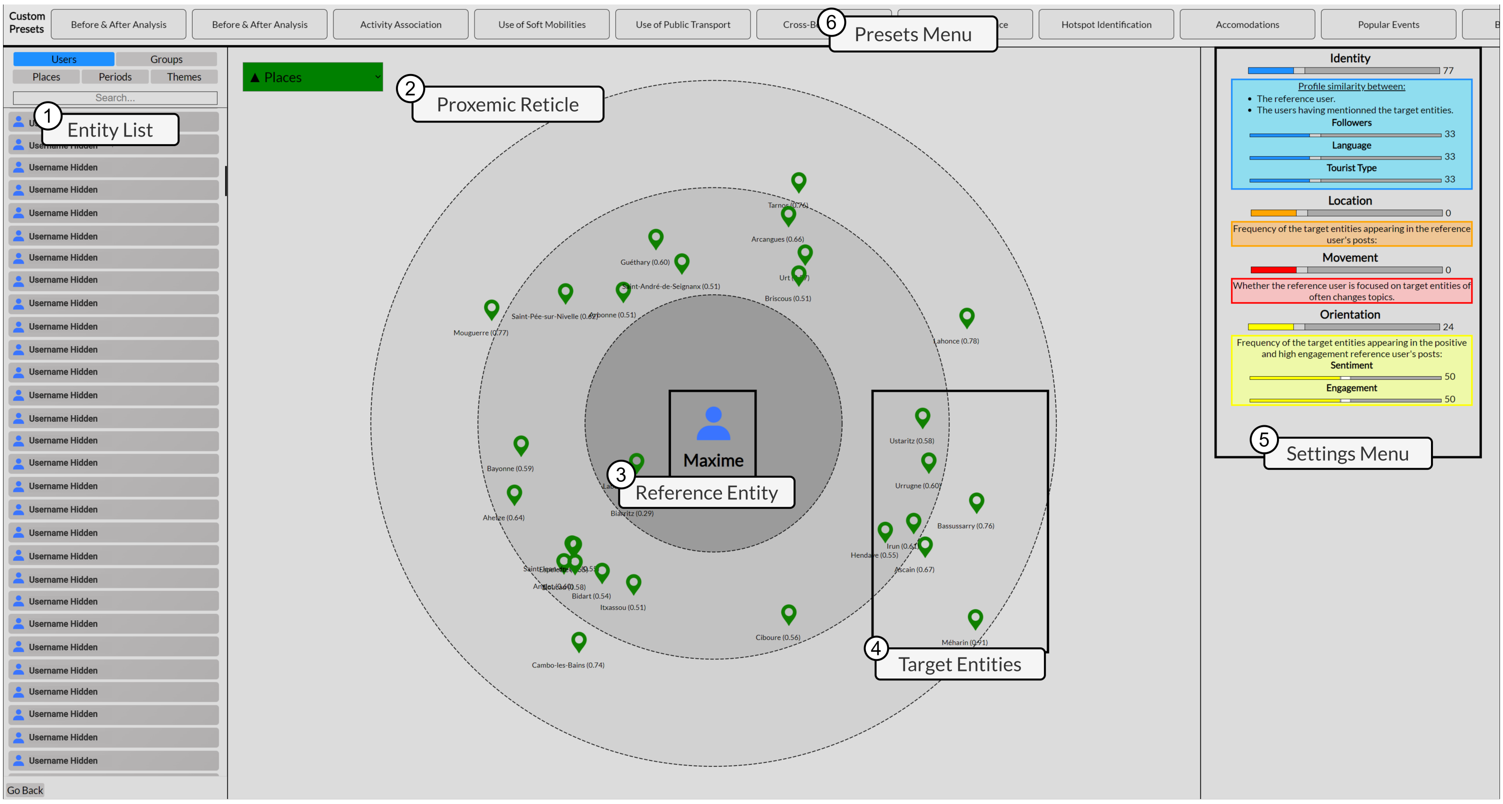
(3) The currently selected *Reference Entity*. We selected the user *Maxime* as a reference by dragging and dropping it from the *Entity List* to the center of the *Proxemic Reticle*.

(4) The currently compared *Target Entities*. As mentioned previously, the type of target entities is selected using the menu at the top left of the proxemic reticle. Here, we observe the affinity of the user *Maxime* with various municipalities in the region.

(5) A *Settings Menu* panel for customizing distance calculations based on the coefficients discussed in Chapter 4. Users can tweak formula settings through sliders, such as attributing greater

importance to positive or highly engaged tweets. Visual results are updated dynamically without requiring reloading. As mentioned earlier, any change is dynamically reflected in the *Proxemic Reticle* without requiring a reload.

(6) A *Presets Menu* panel at the top offers templates of pre-defined distance settings tailored to specific domain requirements. As manipulating proxemic dimensions can be challenging for novice users without training, this allows them to use this view more easily. For instance, in the tourism domain, the *Accommodation* template restricts the display to accommodation-related thematic concepts.



Profile similarity between:

- The reference user.
- The users having mentionned the target entities.

Frequency of the target entities appearing in the reference user's posts:

Whether the reference user is focused on target entities of often changes topics.

Frequency of the target entities appearing in the positive and high engagement reference user's posts:

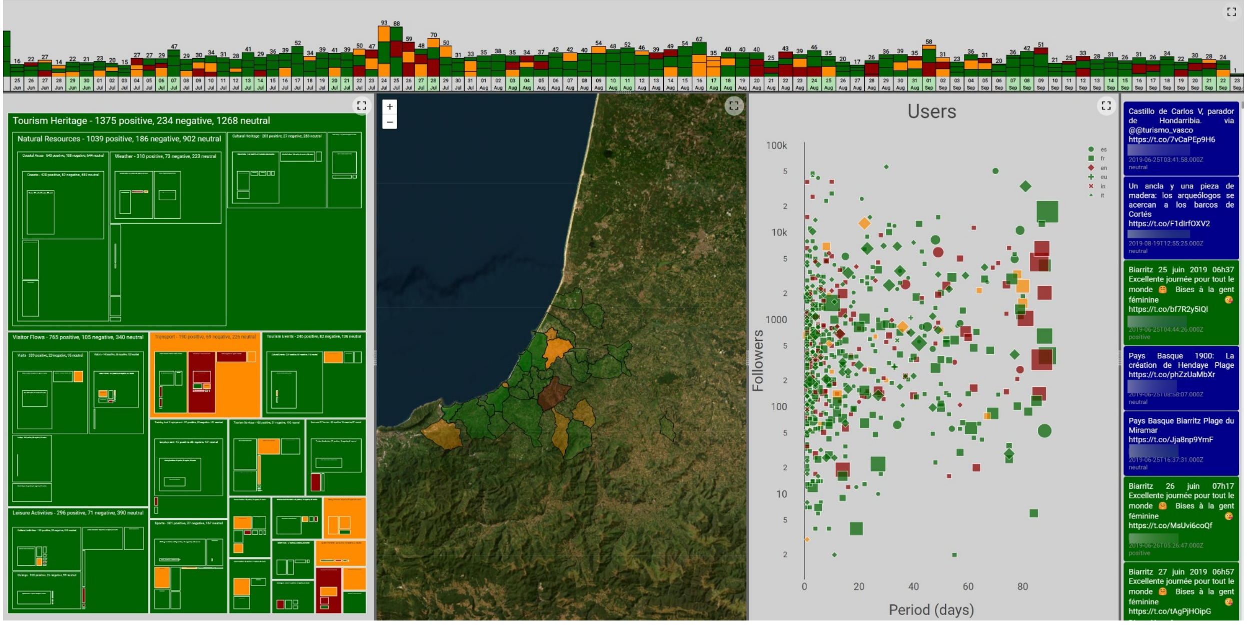
Figure 5.13: Screenshot of the *Proxemics* View

This *proxemics* view is particularly useful for several reasons. The most important one is that it considers several criteria. Unlike the other views of *TextBI*, which focus on specific indicators (e.g., frequency, associations, etc.), the *proxemics* view allows the combination of several indicators in the same visual to address complex end-user requirements. Through proxemic dimension modulation and choice of reference and target entities, it also gives the end user greater control over the visual that will be produced compared to the other views.

5.4.5 Enrichment Overlays

Now, we will demonstrate the two overlays we have designed for the dashboard. Overlays enhance visualization and analysis by providing additional layers of information, enriching the main visuals for specific requirements. They help in better understanding the correlation between different dimensions (for example: sentimental and thematic) and therefore help in decision-making. As a reminder, we have currently implemented two types of overlays: (1) sentiment and (2) engagement.

Figure 5.14 presents the *Frequency* view with the *Sentiment Overlay* enabled. We can observe that most touristic concepts tend to be associated with positive sentiments (*green*), but some, like *Transport* or *Ecology*, are more mixed (*orange*) and even contains negative child concepts (*red*).


Tourism Heritage - 1375 positive, 234 negative, 1268 neutral
Natural Resources - 1039 positive, 186 negative, 902 neutral
Cultural Heritage - 203 positive, 27 negative, 283 neutral
Coastal Areas - 645 positive, 108 negative, 644 neutral
Weather - 310 positive, 73 negative, 223 neutral
Coasts - 420 positive, 82 negative, 489 neutral
Visitor Flows - 765 positive, 105 negative, 340 neutral
Transport - 190 positive, 69 negative, 226 neutral
Tourism Events - 246 positive, 82 negative, 136 neutral
Visits - 539 positive, 23 negative, 95 neutral
Leisure Activities - 296 positive, 71 negative, 390 neutral
Users
Followers
Period (days)
100k
10k
1000
100
10
0
20
40
60
80
es
fr
en
eu
in
it

Castillo de Carlos V, parador de Hondarribia. via @@turismo_vasco https://t.co/7vCaPEp9H6
[redacted]
2019-06-25T03:41:58.000Z
neutral

Un ancla y una pieza de madera: los arqueólogos se acercan a los barcos de Cortés https://t.co/F1dIrfOXV2
[redacted]
2019-08-19T12:55:25.000Z
neutral

Biarritz 25 juin 2019 06h37 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/bf7R2y5IQl
[redacted]
2019-06-25T04:44:26.000Z
positive

Pays Basque 1900: La création de Hendaye Plage https://t.co/phZzUaMbXr
[redacted]
2019-06-25T08:58:07.000Z
neutral

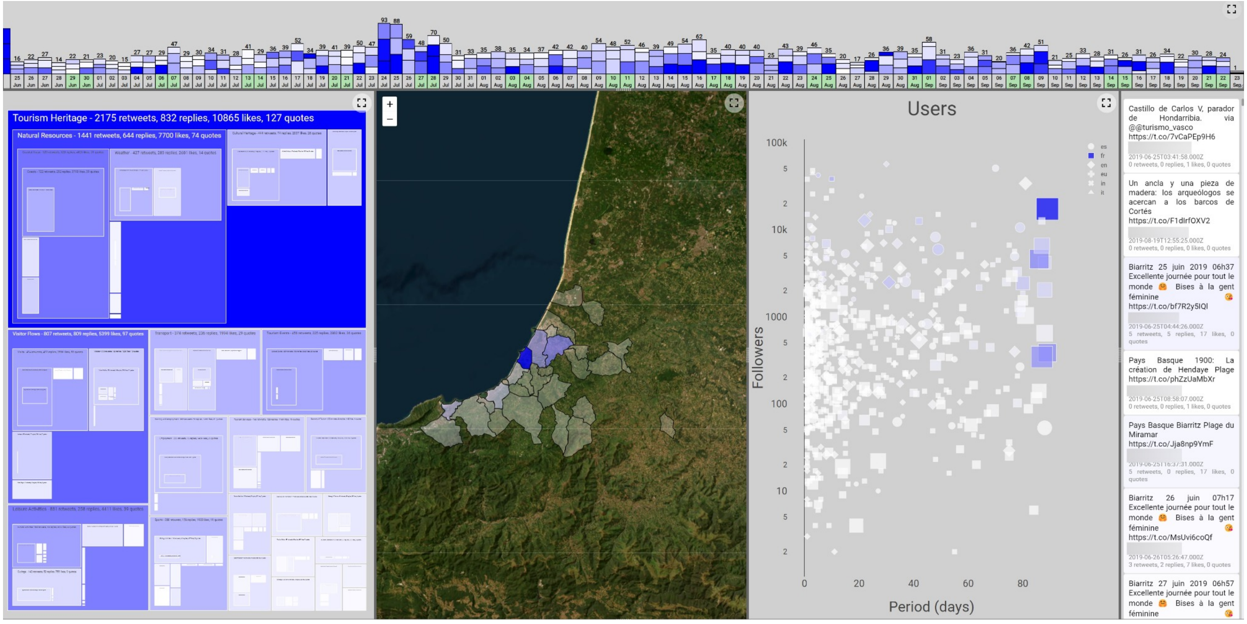
Pays Basque Biarritz Plage du Miramar https://t.co/Jja8np9YmF
[redacted]
2019-06-25T16:37:31.000Z
neutral

Biarritz 26 juin 07h17 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/MsUvi6coQf
[redacted]
2019-06-26T05:26:47.000Z
positive

Biarritz 27 juin 2019 06h57 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/tAgPjHOipG

Figure 5.14: The Frequency View with the Sentiment Overlay Applied

Figure 5.15 presents the *Frequency* view with the *Engagement Overlay* enabled. Here darker blue elements represent entities that are often engaged with by users. For example, we can observe that tweets about tourism heritage-related concepts were retweeted 2,175 times, replied to 832 times, liked 10,865 times, and quoted 127 times.


Tourism Heritage - 2175 retweets, 832 replies, 10865 likes, 127 quotes
Natural Resources - 1441 retweets, 644 replies, 7700 likes, 74 quotes
Cultural Heritage - 444 retweets, 74 replies, 2007 likes, 26 quotes
Weather - 427 retweets, 289 replies, 2601 likes, 14 quotes
Visitor Flows - 807 retweets, 809 replies, 5399 likes, 97 quotes
Transport - 374 retweets, 236 replies, 1994 likes, 29 quotes
Users
Followers
Period (days)
100k
10k
1000
100
10
0
20
40
60
80
es
fr
en
eu
in
it

Castillo de Carlos V, parador de Hondarribia. via @@turismo_vasco https://t.co/7vCaPEp9H6
[redacted]
2019-06-25T03:41:58.000Z
0 retweets, 0 replies, 1 likes, 0 quotes

Un ancla y una pieza de madera: los arqueólogos se acercan a los barcos de Cortés https://t.co/F1dIrfOXV2
[redacted]
2019-08-19T12:55:25.000Z
0 retweets, 0 replies, 0 likes, 0 quotes

Biarritz 25 juin 2019 06h37 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/bf7R2y5IQl
[redacted]
2019-06-25T04:44:26.000Z
5 retweets, 5 replies, 17 likes, 0 quotes

Pays Basque 1900: La création de Hendaye Plage https://t.co/phZzUaMbXr
[redacted]
2019-06-25T08:58:07.000Z
0 retweets, 0 replies, 1 likes, 0 quotes

Pays Basque Biarritz Plage du Miramar https://t.co/Jja8np9YmF
[redacted]
2019-06-25T16:37:31.000Z
5 retweets, 0 replies, 17 likes, 0 quotes

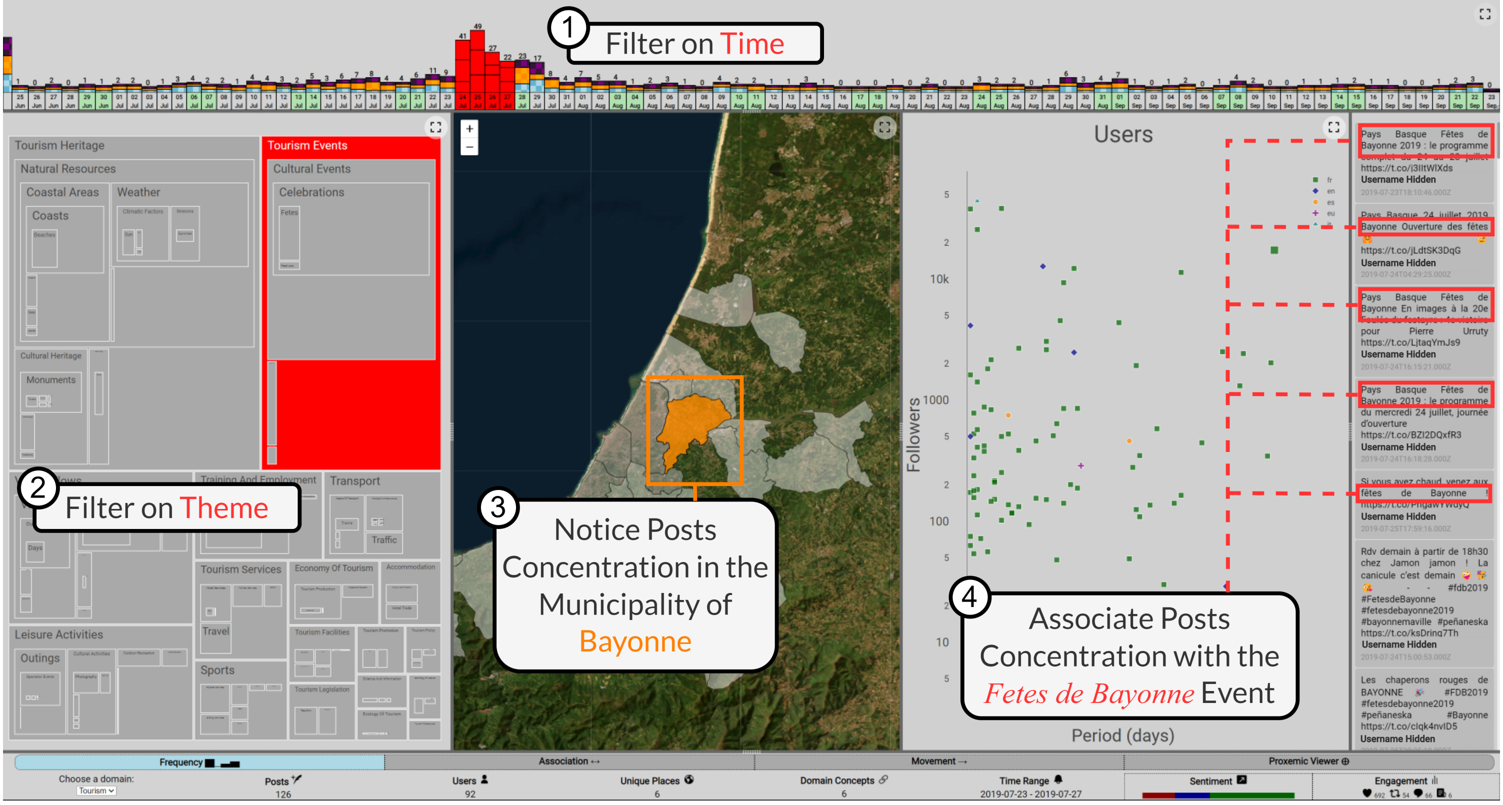
Biarritz 26 juin 07h17 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/MsUvi6coQf
[redacted]
2019-06-26T05:26:47.000Z
3 retweets, 2 replies, 7 likes, 0 quotes

Biarritz 27 juin 2019 06h57 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘

Figure 5.15: The Frequency View with the Engagement Overlay Applied

5.4.6 Interactions and Synchronised Visuals

The dashboard is interactive and features various types of interactions commonly found in BI tools. Additionally, all the visuals are synchronized and updated dynamically without requiring a reload.



Pays Basque Fêtes de Bayonne 2019 : le programme complet du 24 au 28 juillet https://t.co/j3IltWlXds
Username Hidden
2019-07-23T18:10:46.000Z

Pays Basque 24 juillet 2019 Bayonne Ouverture des fêtes 🤗 😉 https://t.co/jLdtSK3DqG
Username Hidden
2019-07-24T04:29:25.000Z

Pays Basque Fêtes de Bayonne En images à la 20e Foulée du festayre : 4e victoire pour Pierre Urruty https://t.co/LjtaqYmJs9
Username Hidden
2019-07-24T16:15:21.000Z

Pays Basque Fêtes de Bayonne 2019 : le programme du mercredi 24 juillet, journée d'ouverture https://t.co/BZI2DQxfR3
Username Hidden
2019-07-24T16:18:28.000Z

Si vous avez chaud venez aux fêtes de Bayonne ! https://t.co/PhgawYWdyQ
Username Hidden
2019-07-25T17:59:16.000Z

Rdv demain à partir de 18h30 chez Jamon jamon ! La canicule c'est demain 🤪 🥳 😘 - - #fdb2019 #FetesdeBayonne #fetesdebayonne2019 #bayonnemaville #peñaneska https://t.co/ksDring7Th
Username Hidden
2019-07-24T15:00:53.000Z

Les chaperons rouges de BAYONNE 🎉 #FDB2019 #fetesdebayonne2019 #peñaneska #Bayonne https://t.co/clqk4nvlD5
Username Hidden

Figure 5.16: The Frequency View After Applying a Tempo-Thematic Filter

Dashboard users can leverage *combined filtering* to restrict the amount of data being presented and to carry out finer analyses. Consider the example depicted in Figure 5.16.

(1) The user selects the time range of July 24th to 27th in the *Timeline*.

(2) The *Thematic Map* updates to display a higher concentration of the *Celebration* thematic concept.

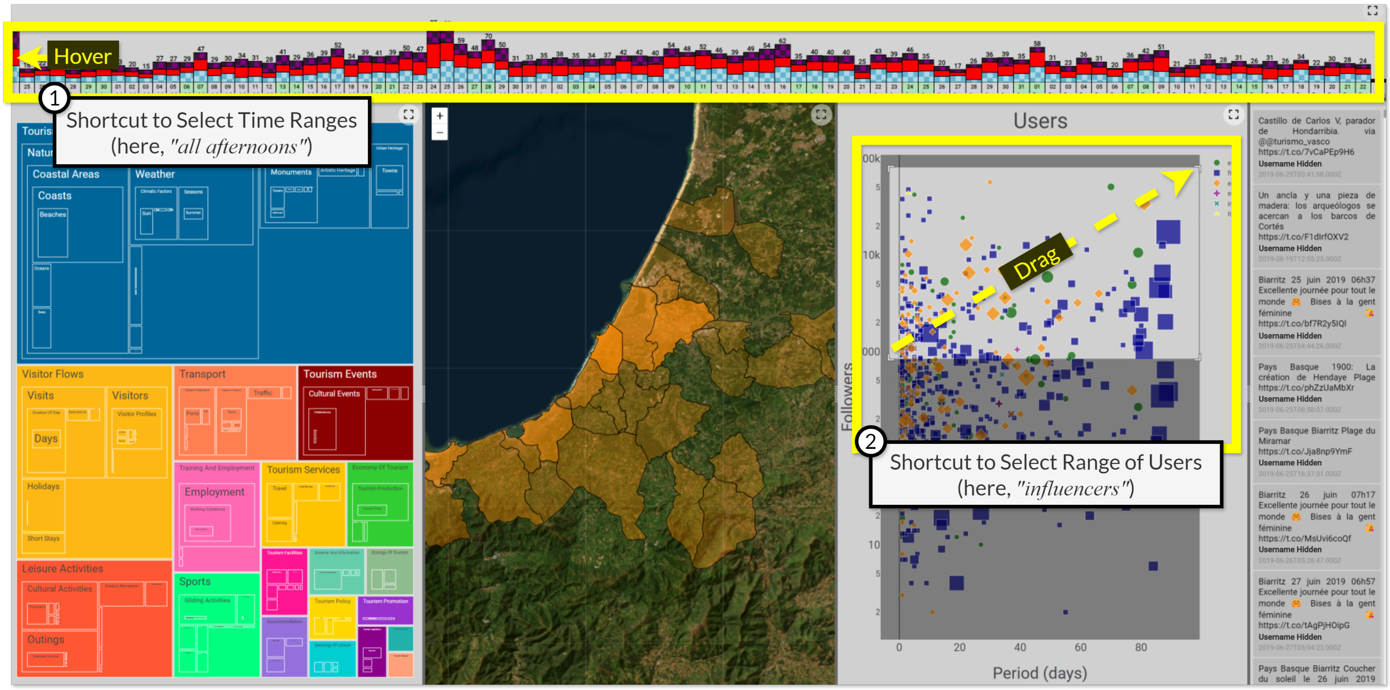
(3) Further clicking on the *Celebration* concept causes the *Spatial Map* to highlight a hotspot in the municipality of *Bayonne*.

(4) By examining the *Posts List*, the user can observe that this coincides with the timing of the *Fêtes de Bayonne* event, a local event that attracts over a million attendees.

Various shortcuts are available to make manipulating the dashboard easier (see Figure 5.17).

(1) To select day ranges over a long period in the *Timeline*, the user can hover over a root element on the left; in this example, we can select all afternoons in only one click.

(2) The *User Map* allows drawing a rectangle to select a wide range of users. This is especially useful for selecting all *influencers*, or all users who tweeted more over a long range of days.



Castillo de Carlos V, parador de Hondarribia. via @@turismo_vasco https://t.co/7vCaPEp9H6
Username Hidden
2019-06-25T03:41:58.000Z

Un ancla y una pieza de madera: los arqueólogos se acercan a los barcos de Cortés https://t.co/F1dlrfOXV2
Username Hidden
2019-08-19T12:55:25.000Z

Biarritz 25 juin 2019 06h37 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/bf7R2y5IQl
Username Hidden
2019-06-25T04:44:26.000Z

Pays Basque 1900: La création de Hendaye Plage https://t.co/phZzUaMbXr
Username Hidden
2019-06-25T08:58:07.000Z

Pays Basque Biarritz Plage du Miramar https://t.co/Jja8np9YmF
Username Hidden
2019-06-25T16:37:31.000Z

Biarritz 26 juin 07h17 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/MsUvi6coQf
Username Hidden
2019-06-26T05:26:47.000Z

Biarritz 27 juin 2019 06h57 Excellente journée pour tout le monde 🤗 Bises à la gent féminine 😘 https://t.co/tAgPjHOipG
Username Hidden
2019-06-27T05:04:23.000Z

Pays Basque Biarritz Coucher du soleil le 26 juin 2019

Figure 5.17: Two Types of Shortcut Interactions Featured in *TextBI*

Finally, the visuals support various granularities to display the data (refer to Figure 5.18). For now, there is no Graphical User Interface (GUI) dedicated to setting the granularity (this is done through a configuration file), but this is planned for the future.

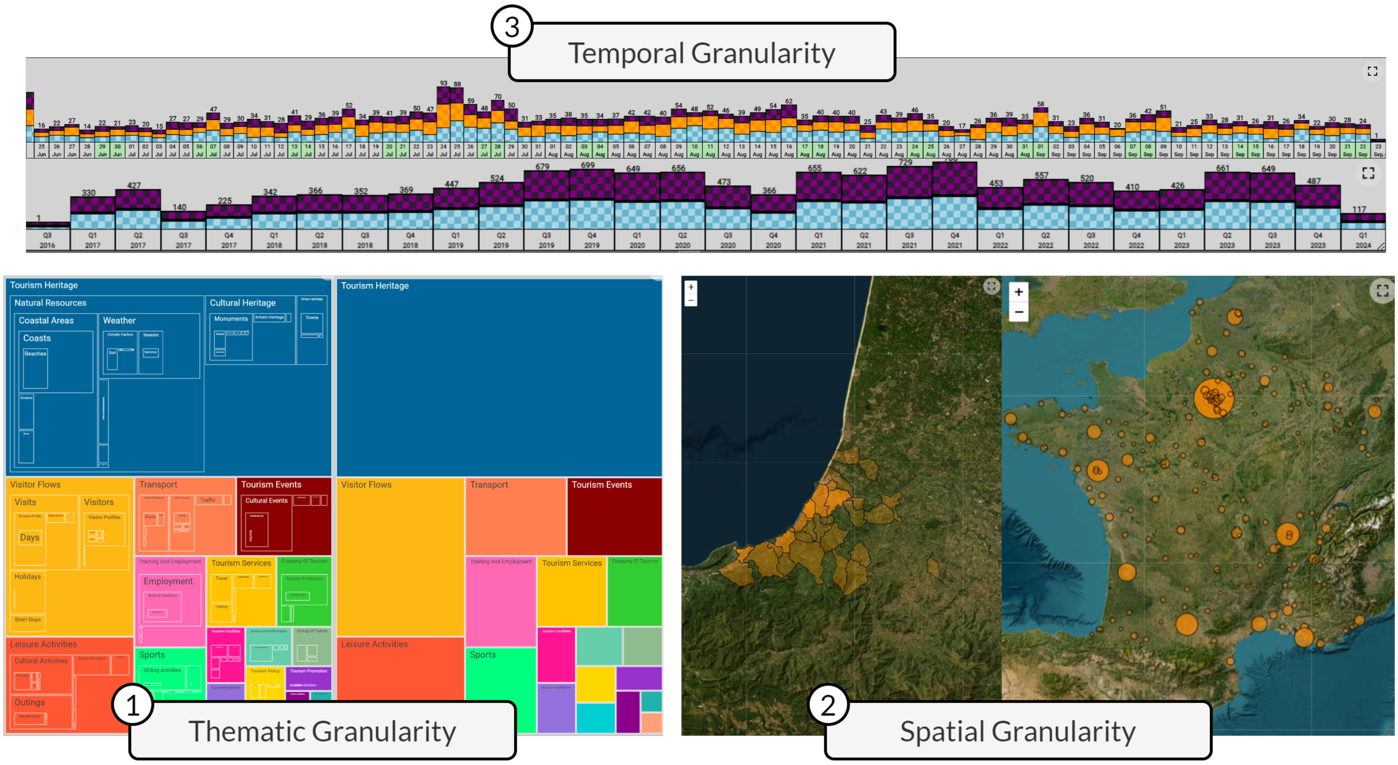


Figure 5.18: Example of Three Visuals with Different Granularity Applied

(1) The *Thematic Map* can be set to display specific levels of the semantic resource used. For

example, only first-level thematic concepts, second-level thematic concepts, etc. This allows for broader or more focused thematic analyses.

(2) The *Spatial Map* can display various spatial granularities. For example, by aggregating the spatial dimension at the POI level, municipality level, region level, country level, etc.

(3) The *Timeline* can be set up to aggregate by different temporal units such as days, weeks, months, semesters, seasons, years, etc.

We now need to assess whether the platform satisfies the requirements of the end users (refer to Figure 5.2). Therefore, we will now qualitatively evaluate this platform with *tourism* stakeholders, who are the main category of end users to whom this platform is dedicated.

5.5 Qualitative Evaluation of the *TextBI* Dashboard

Here, we present a qualitative evaluation of the *TextBI* dashboard. The aim of this evaluation is threefold:

- *Assess and evaluate the dashboard's functionalities* across various dimensions (such as features, design, usability, performance, etc.) with actual end users. Here, we focus on the domain stakeholder category of user which is the main user category of the APs Framework.
- *Validate our working hypothesis*. Namely, an interactive dashboard, centered around four core dimensions: spatial, temporal, thematic, and personal, augmented with data enrichment elements such as sentiment and engagement metrics, and drawing inspiration as well as blending the design principles of BI, GIS, and Linguistic Information Visualization, can offer a user-friendly and adaptable platform for non-computer scientists to easily gain insights into social media annotations across various domains and data sources.
- *Identify areas for improvement*, in terms of both substance (requirements not met by the current dashboard) and form (visual design, usability, etc.).

For this evaluation, we solicited a representative from the domain stakeholder category of end user (refer to Figure 5.2). These are non-computer scientist users seeking to enrich their current analysis processes with social media data. Their goal is to extract insights and valuable indicators from social media to address domain-specific requirements and support their decision-making processes. More precisely, we contacted the director of the *Pau Béarn Pyrénées Tourism Office*[5]. He is a specialist in the domain of tourism and a non-computer scientist user. Currently, most of his decisions are based on the *Flux Vision*[6] tool. It is a platform offered by a French telecom company (*Orange*) that allows the visualization of visitor flows based on their phone location data. The main limitation of this platform is that it focuses on the spatio-temporal dimensions only, and therefore limits the understanding of visitors' behaviors and feelings. The director therefore requires tools that can enrich his analyses with additional dimensions, in particular tools based on social media. We will have him experiment with and evaluate the *TextBI* platform. From now on, we will refer to the director of the tourism office as *the end user*.

[5] https://www.pau.fr/office-du-tourisme

[6] https://www.orange-business.com/fr/solutions/data-intelligence-iot/flux-vision

This section is organized as follows. We start by explaining the evaluation protocol and the evaluation metrics used (see Subsection 5.5.1), present the evaluation results (see Subsection 5.5.2), and finally discuss them (see Subsection 5.5.3).

5.5.1 Evaluation Protocol and Metrics

For this evaluation, we decided to go for a qualitative evaluation approach (Miles and Huberman, 1994, 2003; Ata, 2022). These approaches are frequently used to evaluate software interfaces from the perspective of end users. They have been used to evaluate diverse types of interface in various domains (Kim et al., 2015; Grønli et al., 2014; Hub and Zatloukal, 2008; Tsou and Curran, 2008). Based on the wide array of existing works leveraging qualitative analysis of software (Jooste et al., 2014; Bijarchian and Ali, 2014; Jander et al., 2011; Bastien and Scapin, 1993), we have selected five core criteria to evaluate the *TextBI* platform on:

- *Feel*. This category assesses the overall experience of interacting with the platform. The goal is to understand how the platform makes users feel while they are using it. For example, it's intuitiveness or the feedback it provides.
- *Usability*. This category evaluates how easy and efficient the platform is to use. The focus is on how quickly new users can become proficient and how effectively experienced users can achieve their goals.
- *Design*. This category examines the visual and aesthetic aspects of the platform. The aim is to ensure the platform is visually pleasing and aligns with user expectations for modern tools.
- *Features*. This category looks at the functionality provided by the platform. The emphasis is on ensuring that the platform meets the specific requirements of its intended end users and provides valuable capabilities.
- *Performance*. This category measures how well the platform performs in terms of speed and reliability. The goal is to ensure the platform operates smoothly and efficiently.

For each criterion, we have subdivided them into sub-criteria. Table 5.6 provides an overview of the criteria and sub-criteria that will be used to qualitatively evaluate the *TextBI* dashboard. For each of them, we also present a general *Description* and a *Statement* with which the end user will have to agree or disagree.

We will use a *Likert scale* (Jebb et al., 2021; Joshi et al., 2015) (see Table 5.5) to measure the end user's level of agreement with each statement. The *Likert scale* is a psychometric scale commonly used in surveys. It ranges from 1 to 5, where 1 indicates strong disagreement and 5 indicates strong agreement (Harpe, 2015). This scale allows us to quantify subjective assessments and gather detailed insights into user perceptions and experiences about the *TextBI* platform.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
●○○○○ (1)	●●○○○ (2)	●●●○○ (3)	●●●●○ (4)	●●●●● (5)

Table 5.5: Overview of the Likert Scale

Criteria	ID	Sub-Criteria	Description	Statement
Feel	C1	Intuitiveness	The ease with which users can navigate the platform and predict how interactions will occur.	The platform is intuitive and easy to navigate from the moment you start using it.
	C2	Feedback	The visual responses the platform provides to inform users of the results of their interactions.	The platform provides clear feedback when you perform actions.
Usability	C3	Learnability	How quickly users can learn to navigate the platform and use its features effectively.	Learning how to use the platform is a quick and straightforward process.
	C4	Memorability	The degree to which users can remember how to use the platform after a long period of not using it.	After not using the platform for a while, it is easy to remember how to use it again.
	C5	Ease of use	How straightforward and user-friendly the platform is.	The platform is easy to use, even for someone who may not be very experienced.
	C6	Efficiency	How fast users can complete tasks using the platform after they have become familiar.	Your tasks can be completed efficiently using the platform.
	C7	Recoverability	The platform's ability to facilitate recovery from user mistakes.	If an error occurs in the platform, it is easy to recover and get back on track.
	C8	Error Handling	The platform's capability to handle user errors and offer solutions.	The platform effectively guides you to resolve errors when they happen.
Design	C9	Global Aesthetics	The overall visual appeal of the platform, including styles, consistency, and visual hierarchy.	The overall visual design of the platform is aesthetically pleasing.
	C10	Consistency	The coherence of visual elements such as fonts, colors, and layout throughout the system.	The visual design of the platform maintains consistency across different features and functions.
	C11	Colors	The choice of color to enhance usability, convey status, and guide user attention.	The color scheme of the platform helps in distinguishing different elements clearly.
	C12	Clarity	The legibility and comprehensibility of information presented in the system.	The information presented by the platform is clear and easy to understand.
	C13	Layout	The arrangement of platform elements to allow for efficient interaction and understanding.	The layout of platform elements facilitates easy understanding and interactions.
Features	C14	Completeness	The extent to which the platform provides all the functions required by users.	The platform provides all the functions and features you require.
	C15	Flexibility	The system's ability to accommodate a range of user preferences and workflows.	The platform offers enough flexibility to accommodate your way of working.
	C16	Potential	The capacity of the platform to accommodate potential future analysis requirements	The platform has the potential for future use.
	C17	Usefulness	The usefulness of platform features for the user's intended purposes.	The features provided by the platform are practical and valuable in your daily use.
	C18	Accuracy	The precision with which the platform reflects and executes the desired actions.	The platform accurately performs the tasks and actions you expect.
	C19	Innovativeness	The degree to which the platform introduces novel solutions and design approaches.	The platform is innovative in its design and features compared to what you usually use.
	C20	Compatibility	The ability of the platform to operate with different systems, devices, or standards.	The platform is compatible with other tools and devices you use.
Performance	C21	Reactiveness	The speed and responsiveness of the platform to user inputs.	The response time of the platform is quick enough when you interact with it.
	C22	Stability	The reliability of the platform during extended usage and under different conditions.	The platform is stable and reliable.

Table 5.6: Criteria Used to Qualitatively Evaluate the *TextBI* Dashboard

We aim for an end-user satisfaction rate of 75%, which corresponds to a score of 3.75 out of 5 on the *Likert scale*. This implies an expected 25% disagreement, highlighting areas of the software that may require refinement. It is important to note that the version of the software under evaluation is a prototype, and achieving the higher results at this stage is challenging. We therefore expect significant areas of improvement. The evaluation of the *TextBI* platform with the end user was conducted as follows:

- We presented to the end user an overview of the *TextBI* platform, highlighting its core features and the scope of the dataset loaded into it.
- For each view (e.g., *Frequency*, *Association*, *Movement*, and *Proxemics*) and *Overlays*:
    - We provided a detailed explanation of the view and demonstrated its functionalities to the end user (e.g., visuals and interactions available).
    - The end user was given a few minutes to interact with the view.
    - We presented each statement from Table 5.6 to the end user and asked whether they strongly agreed, agreed, disagreed, strongly disagreed, or were neutral (e.g., neither agreed nor disagreed) regarding the statement. We instructed them to consider only the current view in their answer. The end user was allowed to interact more with the platform if they were unsure.
- We then presented each statement again, this time asking the end user to answer globally, considering the entire platform and not just specific views.

With the evaluation protocol and metrics established, we will now present the results.

5.5.2 Evaluation Results

Table 5.7 presents an aggregated, simplified version of the results for easier interpretation. Here, we have calculated the mean results for each criterion and target (as a reminder, results are expressed between 1 and 5, with 1 corresponding to a strong disagreement and 5 to a strong agreement). The *Details* column in Table 5.7 shows an overview of the individual evaluation for each subcriterion from Table 5.6. For more details, please refer to Appendix D (*Tourism Office* column).

TextBI was evaluated positively for most criteria, with the majority of them being above 4 (agree). Note that the *Error Handling* subcriterion was always assessed as "*Neither Disagree Nor Agree*" (neutral) due to the experimentation time being too short. The *End User's Comments* column in Table 5.7 transcribes the reactions of the users to the corresponding target. These comments have been translated from French and rephrased for better readability.

The main results from Table 5.7 and Appendix D are as follows (we have chosen to focus on negative aspects, as these give us hints about what to improve):

- Most targets received a positive assessment. The end user emphasized the engaging design and reactivity of the dashboard, with immediate feedback, fast visual updates, and mostly intuitive colors.

Target	Criteria	Results	Details	Transcribed End User's Comments
Frequency View	Feel	4.5	4, 5	The view is well presented, engaging and presents interesting information. It is innovative, none of the tools I currently use have these capacities. Prior training may be beneficial to fully understand all aspects. The design is satisfying, though some color adjustments could enhance clarity (e.g., more meaningful colors for thematic concepts in the thematic map). Having a simplified version of each visual, reporting only the most important information could aid understanding. It has significant potential and will be useful in my work, though not for everyday use. It would integrate well with the tools I currently use.
	Usability	4.17	5, 5, 2, 5, 5, 3	
	Design	4.2	5, 5, 4, 2, 5	
	Features	4	1, 5, 5, 2, 5, 5, 5	
	Performance	5	5, 5	
Association View	Feel	4.5	4, 5	This view complements the previous one well. However, learning to use it efficiently is a bit more difficult due to the large amount of information displayed. Simplified visuals are needed so information can be extracted at first glance without having to filter and spend too much time on it. Perhaps two versions of each visual should be offered: the full version and a simplified version (which would be the default), presented like a report. I will probably use this view less than the first one, but it still has frequent use cases.
	Usability	3.5	2, 5, 2, 4, 5, 3	
	Design	4.4	5, 5, 5, 2, 5	
	Features	3.86	1, 5, 5, 2, 4, 5, 5	
	Performance	5	5, 5	
Movement View	Feel	4.5	4, 5	This view fits well into the software by reusing the same design, interactions, and color palette as the previous one. However, it is less innovative. Similar tools already exist in my workflow (Flux Vision), so this one adds less value. Additionally, it is too cluttered; there is too much information displayed, making it difficult to discern arrow directions in the directed graph. I would appreciate a simplified graph showing just the main places where people move from and to. I appreciate the reactiveness of interactions and the instant feedback offered by this view (and the others too).
	Usability	3.17	2, 5, 2, 2, 5, 3	
	Design	3.6	5, 5, 5, 1, 2	
	Features	2.71	1, 3, 3, 1, 4, 2, 5	
	Performance	5	5, 5	
Proxemics View	Feel	5	5, 5	This view is particularly innovative. The setup is easy by simply dragging and dropping entities, and the ability to load presets alleviates the burden of parameterizing the reticle ourselves. Less information is presented at a time compared to previous views, which have multiple visuals, making this view easier for me to manipulate and learn to use. It is innovative, and I see many use cases I could use it for. I particularly liked the ability to create presets as it would allow me to preconfigure indicators I am often interested in (for example to write reports on local tourism) that I would reuse at will.
	Usability	3.83	5, 5, 2, 5, 3, 3	
	Design	4.4	2, 5, 5, 5, 5	
	Features	4	1, 5, 5, 2, 5, 5, 5	
	Performance	5	5, 5	
Overlays	Feel	5	5, 5	Overlays add a new depth of analysis to each view. The colors of the gradient are overall well chosen. However, there are many more dimensions that I would like to analyze besides sentiment and engagement. For example, detailed demographics of visitors evoking a theme or visiting a place (e.g., family visitors, excursionists, inhabitants, etc.) or the influence of recommendations (e.g., by online influencers, by tourism offices themselves, etc.) on visitors' behaviors.
	Usability	4.5	5, 5, 4, 5, 5, 3	
	Design	4	4, 5, 4, 4, 3	
	Features	4.29	1, 5, 5, 4, 5, 5, 5	
	Performance	5	5, 5	
Global	Feel	4.5	4, 5	*TextBI* is a platform I am really interested in. Although it is still a prototype, it offers features that are innovative and could complement my existing tools well. The dynamic interactions and synchronized visuals are impressive, allowing visualization of various relevant dimensions. One thing I note, however, is that the quality of the semantic resource highly influences the quality of the results. The WTO thesaurus is too generic to get a good assessment of local tourism (e.g., local events, monuments, etc.). I would use this platform regularly but not every day.
	Usability	4.17	5, 5, 2, 5, 5, 3	
	Design	4.4	5, 5, 4, 4, 4	
	Features	4.4	2, 5, 5, 4, 5, 5, 5	
	Performance	5	5, 5	

Table 5.7: Aggregated Results of the Qualitative Evaluation Survey of *TextBI* by the Tourism Office, See Appendix D for Detailed Results

- Usability was deemed satisfying, except clarity in some visuals that were considered to present too much information, making it difficult to gain insights rapidly. This issue was primarily present in views featuring graphs (e.g., *Association* and *Movement* views).

- The *Movement* view received significantly worse feedback as it was deemed not innovative enough compared to existing tools like *Flux Vision*, which already provide similar features (but based on telecom data, not social media). Consequently, it offers little added value in terms of features and would not be used regularly.
- The *Proxemics* view was well received with the end user discovering a type of visualization he had never seen before. He particularly appreciated the ability to build custom, multi-criteria indicators and the ability to save them as presets for later use. Indeed, they often have to write reports about the state of local tourism and recurrently have to analyze the same indicators but with updated data (e.g., from the past semester, etc.).
- For the other views, the end user indicated they would use them regularly but not daily, as the current platform does not meet all of their requirements (completeness). For example, the ability to analyze the impact of influencers' suggestions or the tourism office's own suggestions on visitors' behaviors is not covered. The categorization of users by language is too broad; they would like a categorization by type of visitors (e.g., visitors coming to see family, excursionists from nearby municipalities, inhabitants, etc.).
- The end user stressed the importance of the semantic resource used. He deemed the *Thesaurus on Tourism and Leisure Activities of World Tourism Organization* (World Tourism Organization, 2002) as too generic and high-level, making it difficult to extract insights from local tourism which has many particular features (e.g., local events and food).
- Finally, both overlays were received very positively by the end user, who found both of them useful, especially the sentiment overlay, which allows the assessment of visitors' satisfaction regarding touristic cities, touristic themes, and time periods.

Figure 5.19 presents the results from Table 5.7 aggregated by view, displayed as a *Kiviat* diagram (also called radar chart) with equal weighting assigned to all criteria.

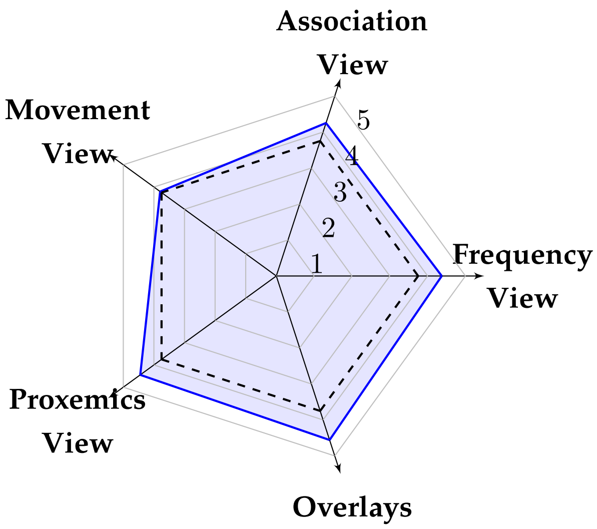


Figure 5.19: Evaluation Results by View

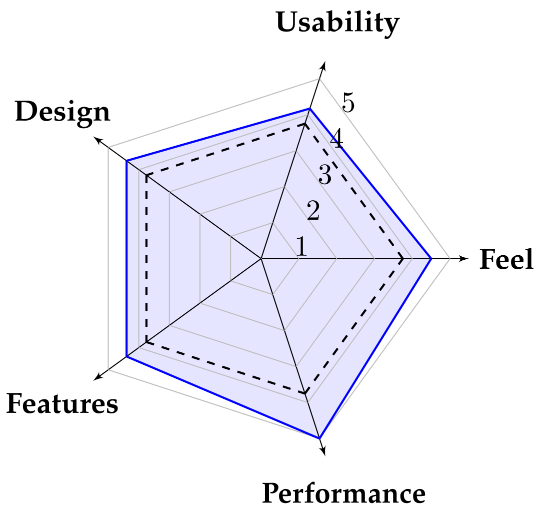


Figure 5.20: Evaluation Results by Criterion

Compared to the expected satisfaction ratio of 75% (equivalent to 3.75 on the *Likert scale*, indicated by the dashed black line in Figure 5.19), all views meet this threshold. However, it is notable that the *Movement* view scores significantly lower than the others. Conversely, Figure 5.20 illustrates the global results, with all criteria exceeding the expected satisfaction ratio. Usability, while still above the threshold, is marginally lower than the other criteria due to the previously discussed issues. Performance is rated exceptionally high (5), likely attributable to the relatively small dataset utilized. To address this, a subsequent experiment with a larger dataset is proposed in Chapter 5. We will now discuss these results and their broader impact on the *TextBI* platform.

5.5.3 Discussion and Limitations

First of all, we have to acknowledge that some criticisms formulated by the end user are not directly related to the *TextBI* dashboard but rather to the specific configuration we used for the experiment. For example, the thesaurus of tourism used as the semantic resource is not bound to the dashboard; a more fine-grained and customized thesaurus could have been used. However, this stresses the importance for the end users of investing time in building a semantic resource as comprehensive as possible to their domain of interest.

Similarly, the user groups we have used (e.g., sorting users by nationality) are not inherent to the dashboard but rather the way the model was instantiated. With more advanced user categorization techniques, we could have defined other groups, and the dashboard would have supported it without requiring significant redesign.

The most recurrent criticism was the fact that the graphs were displaying too much information. It appears therefore necessary to provide two modes for each of them: a complete one like the one we currently have and a simplified one that focuses on the most important information, allowing the stakeholders to immediately view the key insights without spending too much time filtering.

The *Movement* view is, for now, too conventional and offers little value considering many tools propose similar views. We would need to redesign it and propose more original visualizations that are easier to analyze. To this purpose, it is important to increase the dashboard's modularity as new visuals corresponding to new dimensions are necessary to increase the coverage of requirements from end users without cluttering the view too much. We will discuss this further later in the perspectives. Overall, the results of the qualitative evaluation were positive (e.g., globally all criteria were evaluated between 4 and 5, which corresponds to a normal to strong agreement with each statement, see Table 5.7).

The dashboard as a whole is an innovative tool but we have to acknowledge some limitations regarding this experiment. The dataset loaded consisted of ≈ 3000 tweets and was therefore not that massive. This might have biased the *Performance* criteria as the dashboard may have behaved differently with a massive dataset (e.g., longer loading times), we will experiment with a larger dataset in the next chapter (Chapter 6) to observe if the dashboard still perform well. Additionally, due to time constraints, the end users' interaction with the platform was significantly limited.

5.6 Summary and Perspectives

In this chapter, we introduced a novel dashboard named *TextBI* (Contribution 4), designed to facilitate the visualization of social media analyses (e.g., annotations and indicators) for two categories of users: domain stakeholders in various domains of application (non-computer scientists) and NLP researchers (computer scientists). Figure 5.21 presents a visual summary of the dashboard.

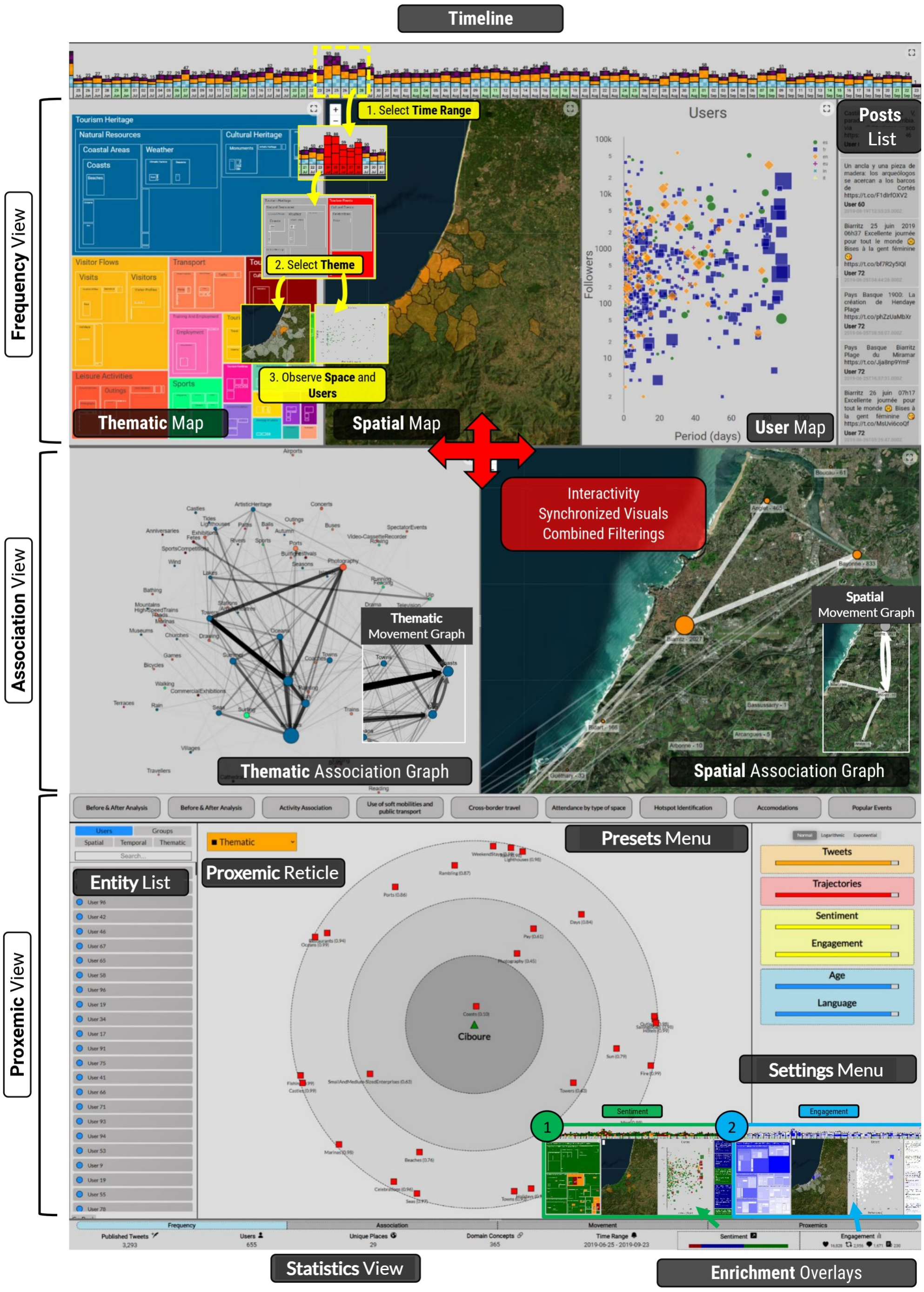


Figure 5.21: Visual Summary of the *TextBI*'s Universal Dashboard

Our literature review suggests that existing solutions may have limitations in several areas. Some are overly focused on single domains with narrow analytical dimensions, such as domain-specific dashboards. Others may not be fully adapted to the analyses required for social media data, particularly in terms of analyzing associations and trajectories, as seen with some BI tools. Additionally, certain solutions appear too complex for users without a background in computer science, such as GIS, some Linguistic Information Visualizations, and computer scientist-oriented visualization libraries. Furthermore, there does not appear to be a current solution that provides an easy-to-use method for setting up and visualizing multi-criteria, domain-adaptable indicators for social media, which is a requirement of domain stakeholders.

Therefore, we propose a generic, domain-adaptable dashboard, *TextBI* (see Figure 5.21). This dashboard's main originality lies in its capability to work with social media corpora from various domains, provided they conform to the data model presented in Chapter 4. It takes as input (1) any semantic domain description along with (2) a social media corpus processed by the APs Framework. Positioned within the research fields of *Visualization*, *Human-Computer Interaction (HCI)*, and *Interactive Systems and Tools*, it aims to address the challenge of presenting social media analyses and indicators across various domains in an accessible manner for non-computer scientist users.

TextBI focuses on four main dimensions: spatial, temporal, thematic, and personal, offering various indicators (e.g., frequency, association, movement, multi-criteria with *proxemics*). It also supports additional enrichment data, such as sentiment and engagement overlays. *TextBI* provides extensive interactivity, including combined filtering, visual synchronization, aggregation, and more. These interactions are inspired by BI tools.

The dashboard was qualitatively evaluated by a domain stakeholder from a tourism office. A qualitative evaluation approach was used, applying various criteria such as *Feel*, *Usability*, *Design*, *Features*, and *Performance*. Overall, the results were positive, and the platform has proven to be interesting for the stakeholders, innovative, had high potential to be integrated into their current analysis process, and was mostly intuitive to use. However, weaknesses were pointed out, such as some views presenting too much information and necessitating simplified versions, as well as the necessity for some prior training on the platform to fully understand everything. The end user also stressed the importance of having a good quality, fine-grained semantic resource as a backbone.

Before industrializing the platform for wider use, various enhancements and additional experiments are essential. These perspectives will be developed further in Section 7.2, but here we provide a brief overview.

Firstly, on the experimental side, it is required to test the platform with massive datasets to observe its performance at scale with datasets in the order of millions of posts, and potentially heterogeneous, meaning sourced from several social media platforms simultaneously.

A thorough review of the color schemes used in the dashboard is crucial, especially regarding map gradients, timeline visuals, and the coloring of proxemic entities. User feedback suggests these features are not always intuitively understood, underscoring the need for enhancements. Research on how colors affect data perception and the rationale behind color palette selections (Ahmad et al., 2021; Szafir, 2017) could provide valuable insights for adopting more intuitive color schemes in *TextBI*.

Furthermore, we aim to make the dashboard customizable by end users. This could be done by having a library of visualization modules (e.g., thematic maps, timelines, association graphs, etc.) that can be dragged onto the main screen to build personalized views. It would also facilitate the creation of new visualization modules. For instance, we are exploring innovative visuals for trajectories, such as metro map-like visualizations (Jacobsen et al., 2020). The motivation behind this perspective is that not all application domains necessarily require all the visualizations.

In particular domains, some dimensions may be more important than others. For example, in the domain of urban planning, the spatial dimension (mapping of resources or infrastructure) and temporal dimension (changes over time) are likely to be more important than the personal dimension, as reflected by existing works in this domain (Isinkaralar, 2023).

Additionally, there is a significant demand for supporting live data to enable real-time tracking of theme evolution and geographical distribution during major events (e.g., elections, sports competitions). This introduces several real-time related complexities (e.g., spatio-temporal computation delays, scalability, real-time implementation of machine-learning algorithms, etc.) (Mehmood and Anees, 2020), as it would necessitate the unification of all APs Framework phases within a single, real-time platform linked with a live database like *InfluxDB* (Ahmad and Ansari, 2017), a feature not included in this thesis due to its engineering complexity and time requirements.

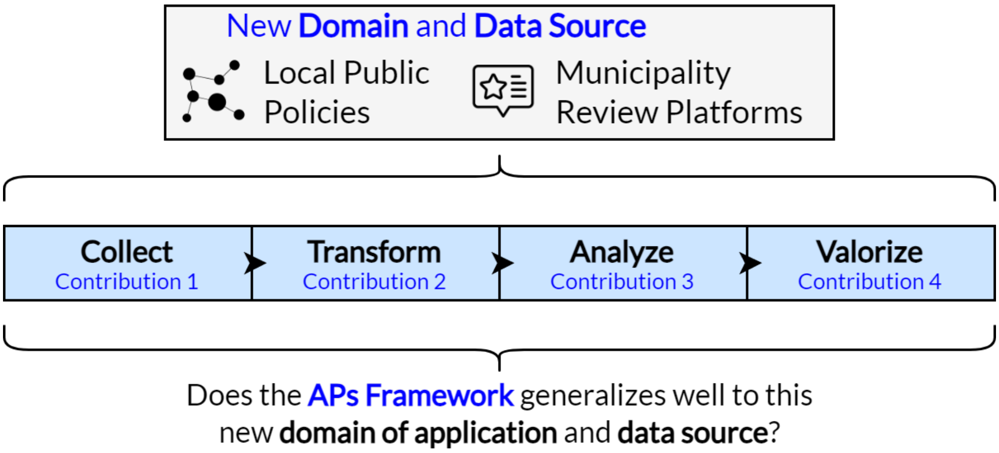
Now, in an effort to showcase the framework's genericity, we will experiment with each of its phases on another data source and domain of application.

Chapter 6

Generalization to Another Data Source and Domain of Application: Municipality Review Platforms and Local Public Policies

"The measure of intelligence is the ability to change."
— Albert Einstein, German Theoretical Physicist

In this chapter, we present a new experiment of the APs Framework (all phases in Figure 1.5) and our proposals (introduced in Chapter 2, Chapter 3, Chapter 4, and Chapter 5) to demonstrate their generalizability to a different domain of application and data source.


New Domain and Data Source
Local Public Policies
Municipality Review Platforms
Collect
Contribution 1
Transform
Contribution 2
Analyze
Contribution 3
Valorize
Contribution 4
Does the APs Framework generalizes well to this new domain of application and data source?

The objective is to demonstrate the genericity of the framework and associated proposals, namely:

- *Domain Genericity*: they support any domain of application expressed using a semantic vocabulary, such as a *dictionary*, *thesaurus*, or *ontology*.
- *Source Genericity*: they are compatible with various post-based social media sources.

In the previous chapter, we experimented with the domain of tourism and the social media platform X/Twitter. Here, we pivot to a new domain, the domain of local public policies, and a new data source: municipality review platforms (namely, *bien-dans-ma-ville.fr*[1]).

We start with a brief introduction highlighting the importance of social media in assessing the perception and impact of public policies in *France* (see Section 6.1), especially for stakeholders at the local level like regional, departmental councils, urban communities, and at the national level. We also introduce the semantic vocabulary we use, a newly built thesaurus of local public policies (see Section 6.2). We then proceed to show the applicability of the domain-independent, iterative data collection methodology to build a corpus of municipality reviews (Contribution 1, see Section 6.3). Following this, we move to the transformation phase and instantiate the APs Trajectory Model using the collected data (Contribution 2, see Section 6.4). After that, we will use the instantiated model along with the redefinition of *proxemics* and proxemic similarity toolkit to generate indicators useful in the domain of local public policies (Contribution 3, see Section 6.5). We used requirements expressed by researchers in local public policies of the OPTIMA research chair[2]. Here the goal is to determine whether the formula and indicators can adapt and produce indicators relevant to this new domain. The last step (Contribution 4, see Section 6.6) is to use the *TextBI* dashboard with this new data source and domain to confirm its genericity. Finally (see Section 6.7), we conclude by discussing the feedback learned from this experiment to pinpoint areas where the framework requires refinement.

6.1 Introduction: The Domain of Local Public Policies

The evaluation of local public policies is increasingly recognized as crucial by governmental bodies, policymakers, academic scholars, and the broader civil society (Ata and Carassus, 2023; Carassus, 2020; Algan et al., 2020). The task of assessing local public policies presents significant challenges, with the efficacy of governmental actions becoming a central topic in societal discourse. Through the analysis of data produced by citizen engagement, it is possible to assess the effectiveness of services provided by local governments. Such evaluations are performed in alignment with the strategic segmentation of policy initiatives (Thoenig, 2010).

Citizens possess various avenues to voice their perspectives and recommendations on various subjects. Various questions arise: How can local governments collect and systematically analyze this feedback to enhance their understanding of the requirements and opinions of citizens, thereby refining their policy measures? Furthermore, what methodologies can be applied to effectively analyze textual feedback related to local public policies?

There is a growing consensus that leveraging large datasets can substantially support the decision-making processes of local stakeholders (e.g., municipal or regional councils) (Kinra et al., 2020; Höchtl et al., 2016), aiding in the enhancement, creation, and strategic planning of local public policies. This involves a detailed examination of the data to better comprehend both the practices and requirements of citizens. Such insights are particularly valuable for authorities striving to meet the expectations and aspirations of their constituents. Various user-generated data sources can be utilized to obtain feedback from citizens, for example, social media (Buccafurri et al., 2012), surveys

[1] https://www.bien-dans-ma-ville.fr

[2] https://optima.univ-pau.fr

(Andrews et al., 2004), public opinion polls (Fishkin, 2003), and various dedicated platforms.

Among these dedicated platforms, municipality review platforms have emerged as a way for citizens to voice their opinions about local public policies at the municipality level. Such platforms include, for *France*: *bien-dans-ma-ville.fr*[3], *ville-ideale.fr*[4], *villesavivre.fr*[5], or *monaviscitoyen.fr*[6].

For this experiment, we have chosen to use municipality review platforms in contrast to social media for several reasons:

- *Termination of Academic Access to X/Twitter API*. The academic API (Application Programming Interface) used to collect tweets from X/Twitter was discontinued in April 2023. Only the paid enterprise plan remains available. Therefore, we could not use it to facilitate the collection process.

- *More Focused*. They are more focused on our domain of interest (local public policies) and therefore less noisy than regular social media, potentially leading to more in-depth analyses and easier framework setup (e.g., less filtering needed).

- *Other Source*. By using a different type of data source, we can validate that the APs Framework can work with another category of user-generated content based on a post system, even if it is not a regular social media.

- *Share Advantages of Regular Social Media*. These platforms share the same advantages of regular social media, such as *ease of access* or *affordability*. However, as mentioned earlier, they are more focused, smaller scale, and not as generalist as regular social media.

Note that in this experiment, we have decided to focus on *France* and opinions expressed in the French language. The reason is that we have already experimented with a multilingual resource in previous experiments on the domain of tourism. Here, we want to focus solely on the new domain and data source. We will explain in Section 6.3 how we have collected data from these platforms using our collection methodology. But firstly (Section 6.2), let's introduce the semantic resource we build to describe the domain of local public policies.

6.2 Creating a Semantic Resource for Local Public Policies

Before proceeding with the collection process, it is necessary to define the semantic resource we use to describe the domain of local public policies. The APs Framework (refer to Figure 1.2) requires to have the semantics of the domain expressed as a dictionary, a thesaurus, or an ontology.

Semantic resources in the field of local public policies in *France* are rare. The closest resources we have found is the *Thesaurus of Public Information* (*Thésaurus de l'Information Publique*) managed by the French directorate of legal and administrative information (DILA)[7]. It was created to index all the texts, speeches, interviews, press releases, press conferences, etc. listed in the collection of public speeches (*Collection des Discours Publiques*)[8]. It provides a semantic tree covering the various fields of public information and public policies.

[3] https://www.bien-dans-ma-ville.fr

[4] https://www.ville-ideale.fr

[5] https://www.villesavivre.fr

[6] https://www.monaviscitoyen.fr

[7] https://www.data.gouv.fr/fr/datasets/thesaurus-information-publique-vie-publique-fr

[8] https://www.vie-publique.fr/collection-discours-publics

But this resource turned out to be too broad for our annotation requirements. An issue that has arisen previously in the tourism domain (refer to Subsection 5.5.2). Indeed, we are interested in *local public policies*, for example, at the municipal or regional level, not in national policies. To create a suitable thesaurus for this experiment, we worked with colleagues from the OPTIMA research chair (Observatory of Local Management and Management Innovation)[9].

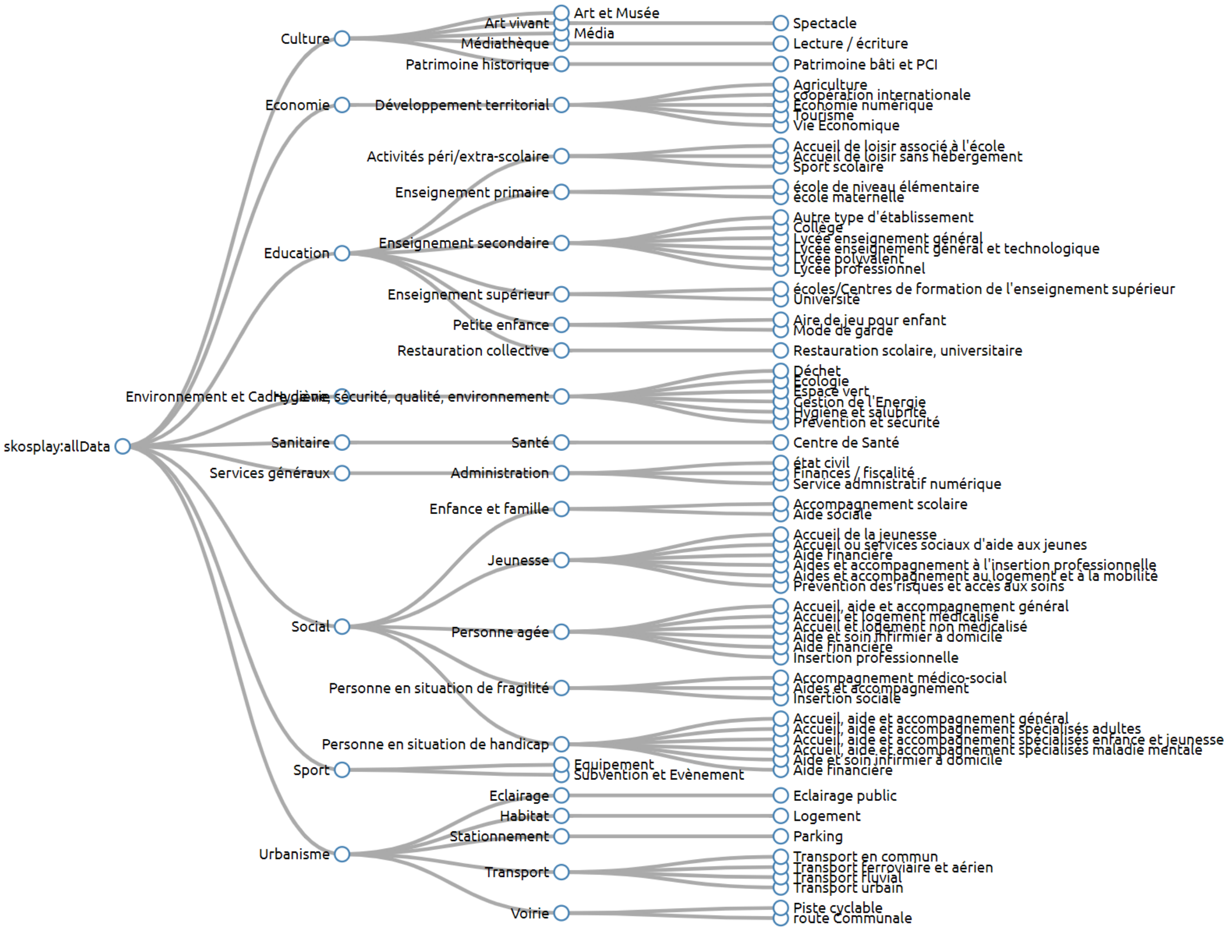

skosplay:allData
Culture
Art et Musée
Art vivant
Média
Médiathèque
Patrimoine historique
Spectacle
Lecture / écriture
Patrimoine bâti et PCI
Economie
Développement territorial
Agriculture
coopération internationale
Economie numérique
Tourisme
Vie Economique
Education
Activités péri/extra-scolaire
Accueil de loisir associé à l'école
Accueil de loisir sans hébergement
Sport scolaire
Enseignement primaire
école de niveau élémentaire
école maternelle
Enseignement secondaire
Autre type d'établissement
Collège
Lycée enseignement général
Lycée enseignement général et technologique
Lycée polyvalent
Lycée professionnel
Enseignement supérieur
écoles/Centres de formation de l'enseignement supérieur
Université
Petite enfance
Aire de jeu pour enfant
Mode de garde
Restauration collective
Restauration scolaire, universitaire
Environnement et Cadre de vie
Hygiène, sécurité, qualité, environnement
Déchet
Ecologie
Espace vert
Gestion de l'Energie
Hygiène et salubrité
Prévention et sécurité
Sanitaire
Santé
Centre de Santé
Services généraux
Administration
état civil
Finances / fiscalité
Service admnistratif numérique
Social
Enfance et famille
Accompagnement scolaire
Aide sociale
Jeunesse
Accueil de la jeunesse
Accueil ou services sociaux d'aide aux jeunes
Aide financière
Aides et accompagnement à l'insertion professionnelle
Aides et accompagnement au logement et à la mobilité
Prévention des risques et accès aux soins
Personne agée
Accueil, aide et accompagnement général
Accueil et logement médicalisé
Accueil et logement non médicalisé
Aide et soin infirmier à domicile
Aide financière
Insertion professionnelle
Personne en situation de fragilité
Accompagnement médico-social
Aides et accompagnement
Insertion sociale
Personne en situation de handicap
Accueil, aide et accompagnement général
Accueil, aide et accompagnement spécialisés adultes
Accueil, aide et accompagnement spécialisés enfance et jeunesse
Accueil, aide et accompagnement spécialisés maladie mentale
Aide et soin infirmier à domicile
Aide financière
Sport
Equipement
Subvention et Evènement
Urbanisme
Eclairage
Eclairage public
Habitat
Logement
Stationnement
Parking
Transport
Transport en commun
Transport ferroviaire et aérien
Transport fluvial
Transport urbain
Voirie
Piste cyclable
route Communale

Figure 6.1: Overview of our Thesaurus of Local Public Policies

The methodology to create the new thesaurus is based on a dual approach, documentary and linguistic. We will not go into details as it is not the main purpose of this chapter but here is an overview:

1. Construction of base corpora (e.g., press corpora, glossaries, etc.).
2. Semantic analysis of selected resources (e.g., frequency of terms, etc.).
3. Creation of a base terminology (e.g., key terms).
4. Design of the thesaurus (using the key terms identified previously, grouping them into various concepts, identifying synonyms, etc.)

To design our thesaurus of local public policies, we have identified nine root concepts. These elements are *Urbanism*, *Education*, *Social*, *Health*, *Culture*, *Sports*, *General Services*, *Environment /*

[9] https://optima.univ-pau.fr

Quality of Life, and *Economy*. We specified the thesaurus in SKOS (*Simple Knowledge Organization System*) language (Miles and Bechhofer, 2009). Figure 6.1 is a graphic representation of our thesaurus (this figure was generated using the SKOS-Play platform[10]).

This resource covers more than 100 domain-specific concepts. We have also defined a set of synonyms for the last sub-concepts. For now the thesaurus is in French only. Below is an example of the *Transport* concept in SKOS format (see Figure 6.2):

```
<skos:Concept rdf:about="http://example.org/Urbanisme/Transport/Transportencommun">
<skos:prefLabel xml:lang="fr">Transport en commun</skos:prefLabel>
<skos:broader rdf:resource="POLICIES://Urbanisme/Transport"/>
    <skos:altLabel xml:lang="fr">Bus</skos:altLabel>
    <skos:altLabel xml:lang="fr">Autocar</skos:altLabel>
    <skos:altLabel xml:lang="fr">Autobus</skos:altLabel>
    <skos:altLabel xml:lang="fr">Co-voiturage</skos:altLabel>
    <skos:altLabel xml:lang="fr">Metro</skos:altLabel>
    <skos:altLabel xml:lang="fr">RER</skos:altLabel>
    <skos:altLabel xml:lang="fr">Tramway</skos:altLabel>
    <skos:altLabel xml:lang="fr">Tram</skos:altLabel>
</skos:Concept>
```

Figure 6.2: Specification of the *Transport* Concept in SKOS Format

We will now explain how we used this resource along with municipality review platforms to build a dataset for use in the APs Framework.

6.3 Applying the Collection Methodology to Local Public Policies

Here, we apply the generic and iterative methodology for constructing thematic datasets (Contribution 1) presented in Chapter 2 to this new application domain (local public policies) and data sources (municipality review platforms). The objective is to demonstrate its genericity and adaptability.

6.3.1 Setup

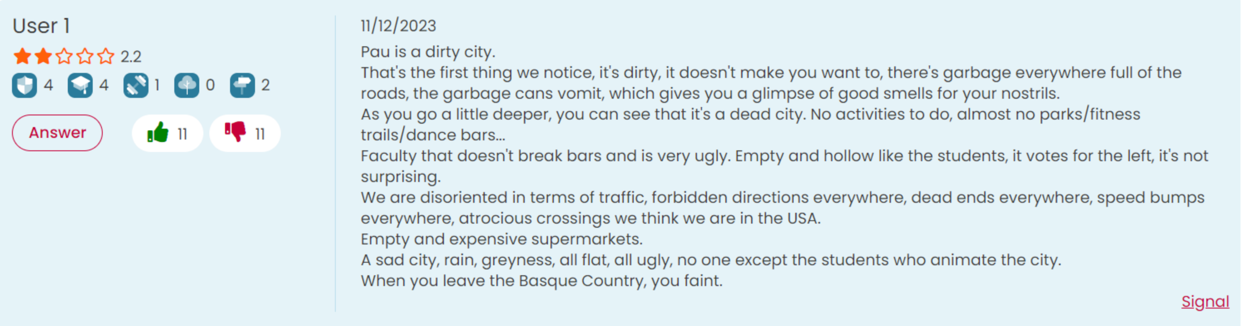
For this experiment, we use the platform *bien-dans-ma-ville.fr*[11], a digital forum that enables residents of various French municipalities to interact, share insights, and voice concerns about urban living and local public policies. This platform serves as a tool for civic engagement. Residents can participate in surveys, contribute to discussions, and provide reviews on local projects and services, making it a valuable resource for gathering data on citizen opinions of local public policies. Here, we will focus on the review aspect of this platform.

Figure 6.3 shows an example of a municipality review extracted from the platform. As we can observe, they can be assimilated to social media posts in their structure, featuring a text that contains spatial and thematic entities, along with metadata capturing sentiment, engagement, and temporal aspects (this will be detailed later, in Figure 6.4). The main difference from social media

[10] https://skos-play.sparna.fr/play

[11] https://www.bien-dans-ma-ville.fr

posts lies in the structure of the text. Municipality reviews tend to be more structured and longer than social media posts, and they generally do not include hashtags or emojis. As we will elaborate further, this characteristic of longer length will be the main limitation of our framework in this context.



11/12/2023

Pau is a dirty city.
That's the first thing we notice, it's dirty, it doesn't make you want to, there's garbage everywhere full of the roads, the garbage cans vomit, which gives you a glimpse of good smells for your nostrils.
As you go a little deeper, you can see that it's a dead city. No activities to do, almost no parks/fitness trails/dance bars...
Faculty that doesn't break bars and is very ugly. Empty and hollow like the students, it votes for the left, it's not surprising.
We are disoriented in terms of traffic, forbidden directions everywhere, dead ends everywhere, speed bumps everywhere, atrocious crossings we think we are in the USA.
Empty and expensive supermarkets.
A sad city, rain, greyness, all flat, all ugly, no one except the students who animate the city.
When you leave the Basque Country, you faint.

Figure 6.3: Example of a Municipality Review on *bien-dans-ma-ville.fr* (translated from French).

We will now demonstrate how we apply our generic data collection methodology to this new data source and domain of application.

6.3.2 Application Process

Table 6.1 shows the application of the generic data collection method to this new domain. For a reminder of the theoretical aspects of the method, please refer to Chapter 2. As this municipality review platform does not provide a dedicated API, we have used web scraping techniques to extract the reviews, more precisely through a *Python* scraper leveraging the *Selenium* (Raghavendra, 2021) headless browser and the *BeautifulSoup* (Richardson, 2007) library.

We proceeded with five refining iterations. All iterations are preceded by a pre-filtering process excluding non-French, very short reviews (less than 10 characters) (to exclude spam), and reviews with a high number of downvotes (more than 50), as we consider those reviews not to be trustworthy (see Table 6.1, *Pre-Filtering*). For each iteration, we have a single filtering flow, as all municipality reviews are geotagged. The domain specialist role active during feedback loops (as a reminder, see Subsection 2.3.4) will be played by a management researcher from the OPTIMA research chair. At each iteration, they are presented with an excerpt of 50 random reviews they have to evaluate (e.g., rate them as suitable or not suitable for analysis) and, in case of issues, propose filter adjustments.

Iteration 1

The first iteration (see Table 6.1, *Iteration 1*) focuses on the *Paris* area (all boroughs) and only recent reviews, less than a year old. We collected 947 reviews, with 56 reviews issued less than a year ago. Among those, 39 contained labels or synonyms of thematic concepts linked to our initial, non-refined thesaurus of local public policies, around 21% of the thesaurus is therefore instantiated with transport-related concepts being highly prevalent. The accuracy (@ 50) of the produced dataset is quite high (0.78) reflecting the relatively high quality of the data source used for this domain, especially in contrast to noisy social media sources.

		Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Pre-Filtering	**Criteria**	**Language**: *French*, **Blacklist**: *Review Length <10 characters, More than 50 downvotes*				
Spatial Filtering	**Criteria**	**Paris** (all boroughs).	**Paris** (all boroughs) and **Top 1 municipality** in each French department.	**Paris** (all boroughs) and **Top 3 municipalities** in each French department.	**Paris** (all boroughs) and **Top 5 municipalities** in each French department.	
	Reviews	947 reviews	9,798 reviews	14,562 reviews	17,146 reviews	17,146 reviews
Temporal Filtering	**Criteria**	Less than **1 year old**	Less than **1 year old**	Less than **3 years old**	Less than **5 years old**	Less than **8 years old**
	Reviews	56 reviews	1,146 reviews	6,330 reviews	13,481 reviews	17,146 reviews
Thematic Filtering	**Criteria**	**Initial** Local Public Policies Thesaurus	**Refined** local public policies Thesaurus - Removed: *Eclairage*, *Tour*, *Subvention* - Added: Synonyms (*subvention sportive*, *taxe foncière*)		**More Refined** local public policies Thesaurus - Restructuration of *Sports* branch - Restructuration of *Culture* branch	
	Reviews	39 reviews	764 reviews	3,977 reviews	8,201 reviews	9,785 reviews
Quantitative Statistics	**Users**	29 users	687 users	3,242 users	6,303 users	7,619 users
	Unique Municipalities	1 municipality	91 municipalities	272 municipalities	454 municipalities	456 municipalities
	Top Municipalities	Paris: 39	Paris: 36 Nantes: 35 Montpellier: 28 Orléans: 26 Avignon: 26 Saint-Brieuc: 21 Nice: 20 Grenoble: 18 Toulon: 17 Bordeaux: 16	Paris: 186 Nantes: 108 Montpellier: 101 Avignon: 78 Bordeaux: 71 Toulouse: 69 Nice: 64 Strasbourg: 61 Grenoble: 58 Brest: 52	Paris: 380 Nantes: 244 Toulouse: 211 Montpellier: 171 Bordeaux: 156 Nice: 108 Rennes: 104 Avignon: 104 Strasbourg: 104 Lille: 98	Paris: 518 Nantes: 305 Toulouse: 270 Bordeaux: 211 Montpellier: 209 Lille: 148 Rennes: 147 Strasbourg: 131 Nice: 125 Avignon: 111
	Unique Concepts	21 (20% of thesaurus)	52 (48% of thesaurus)	63 (60,5% of thesaurus)	65 (62,5 % of thesaurus)	65 (62,5 % of thesaurus)
	Top Concepts	Transport: 10 TransportEnCommun: 8 Logement: 8 Dechet: 7 VieEconomique: 7 PreventionEtSecurite: 6 Education: 4 ArtEtMusee: 4 TrainEtAérien: 4 Culture: 3	PreventionEtSecurite: 164 VieEconomique: 152 Transport: 149 TransportEnCommun: 127 TrainEtAérien: 104 Logement: 93 Culture: 65 Environnement: 60 EspaceVert: 58 Spectacle: 58	VieEconomique: 1,015 PreventionEtSecurite: 779 Transport: 772 TransportEnCommun: 575 TrainEtAérien: 465 Logement: 422 Environnement: 329 Culture: 309 Sport: 277 EspaceVert: 271	VieEconomique: 2,342 Transport: 1680 PreventionEtSecurite: 1,493 TransportEnCommun: 1,150 TrainEtAérien: 841 Logement: 749 Culture: 654 Environnement: 628 Sport: 573 Spectacle: 499	VieEconomique: 2801 Transport: 2091 PreventionEtSecurite: 1685 TransportEnCommun: 1327 TrainEtAérien: 929 Culture: 835 Logement: 830 Environnement: 724 Sport: 665 Spectacle: 597
	Accuracy @ 50	0.78	0.86	0.84	0.85	0.83

Table 6.1: Application of the Data Collection Methodology using our Local Public Policies Dataset Requirements

Iteration 2

The second iteration (see Table 6.1, *Iteration 2*) extends the first one to include, in addition to *Paris* and its boroughs, the most populous municipalities of each metropolitan French department (90 municipalities out of 96 departments, as some municipalities' departments did not have reviews associated). We also refined the thesaurus of local public policies due to some issues identified in the first iteration. For example, some concepts were too broad, like *Éclairage* or *Subvention*, while others, like *Tour*, were causing too much noise (e.g., *une tour*, *le Tour de France*, *faire un tour*). Instead, we added more focused synonyms to existing thematic concepts like *taxe foncière* or *subvention sportive*. All these modifications allowed us to collect 764 reviews belonging to 687 users in 91 municipalities. The accuracy (@ 50) of the dataset produced in this iteration is slightly higher at 0.86 reflecting the impact of the feedback loop (which triggered a thesaurus refinement) on data quality.

Iteration 3

The third iteration (see Table 6.1, *Iteration 3*) is quite similar to the second one, except we extend the spatial footprint of the data to the top 3 most populous municipalities of each French department and extend the reviews collected to the last 3 years. The thesaurus remains unchanged. We therefore collected 3,977 reviews from 3,242 users in 272 municipalities. Interestingly, the most populous municipalities in *France* are not necessarily the ones with the most reviews mentioning local public policies. For example, *Montpellier* and *Nice* are above *Toulouse* despite being smaller municipalities. At this iteration, 63 concepts are instantiated in the thesaurus, which represents around 60.5% of it. The most frequent concepts have changed compared to earlier iterations; now *VieÉconomique* and *PréventionetSécurité* are the most frequent, with 1,015 occurrences and 779 occurrences, respectively. It appears that economic well-being is a more frequent concern in smaller municipalities.

Iteration 4 and 5

Finally, the two latest iterations (see Table 6.1, *Iteration 4* and *Iteration 5*) extend the collection process to the top 5 most populous municipalities in each department (in addition to *Paris*) and review data collected successively in the last 5 years, and in the last 8 years. The thesaurus was redefined again with a restructuring of the *Sport* and *Culture* branches for a more fine-grained analysis of both concept branches. The final dataset (highlighted in green) consists of 9,785 reviews, belonging to 7,619 users. In contrast to the X/Twitter dataset in the tourism domain where each user had an average of around 4.5 tweets, here users have an average of 1.3 reviews so the impact of trajectories should be less significant. Of the 114 thematic concepts of local public policies, 65 (62.5%) are instantiated, and the dataset covers 456 municipalities all around *France*. The most frequent concepts are similar to those in the previous iterations and the accuracy (@ 50) is high at 0.83. We decided to stop the iterations at this stage as we have a high quality (accuracy @ 50 of 0.83) and rather voluminous dataset (more than three times larger than the one used in tourism).

Let's now discuss what we have learned from the application of our methodology to this new domain and data source.

6.3.3 Discussion

This experiment demonstrated that the generic and iterative data collection methodology proposed in the first phase of the APs Framework (Contribution 1) is adaptable for use with various data sources and domains of application. Here, we moved away from building a dataset in the tourism domain using X/Twitter (see Section 2.4) to building a dataset of municipality reviews related to the domain of local public policies. As we have observed, the methodology adapts well to this new domain and data source and allows us to build a dataset vast enough for meaningful analysis.

Overall, it is easier to obtain accurate results using municipality review platforms than regular social media because all reviews are geotagged, making the spatial filtering process easier. However, unlike social media, reviews tend to be longer, which means a given review can be associated with many thematic concepts. This could be a limitation in future phases of the framework. We will elaborate on this later. The temporal filtering is similar to the previous X/Twitter experiment on tourism (e.g., it leverages timestamps).

We will now be moving on to the second step of the APs Framework, data transformation along with this newly collected dataset.

6.4 Transforming Data and Instantiating the Model

We now shift our focus to the data transformation phase (refer to Chapter 3, Contribution 2). To recap, this phase aims to extract structured data from unstructured text. Our primary interest is in determining sentiment polarity at the text level and identifying places and fine-grained thematic concepts at the token level. The extracted data is then used to populate the APs Trajectory Model (refer to Chapter 4, Contribution 3.2) based on a formal redefinition of *proxemics* for social media (Contribution 3.1).

Our framework can accommodate any technique for extracting this knowledge. In Chapter 3, we presented a comparative study of NLP strategies, focusing on deep learning-based techniques for processing social media data in the tourism domain. This study, based on a novel, annotated dataset for the tourism domain derived from X/Twitter explored the most effective strategies based on the number of annotated examples available.

Given the significant time and effort required to develop a novel manually annotated dataset for the local public policies domain (we could not find any existing annotated resource), we opted for alternative techniques to extract the necessary structured knowledge from reviews.

1. For *sentiments*, we exploited the ratings associated with reviews. Unlike typical social media posts, reviews include a numeric star rating, ranging from 1 to 5 stars (see Figure 6.3). Using the star-to-sentiment mapping proposed in Sharma and Dutta (2021), we infer the sentiment expressed in the review: 1 to 2 stars denote a negative sentiment, 4 to 5 stars a positive one, and 3 stars a neutral sentiment.

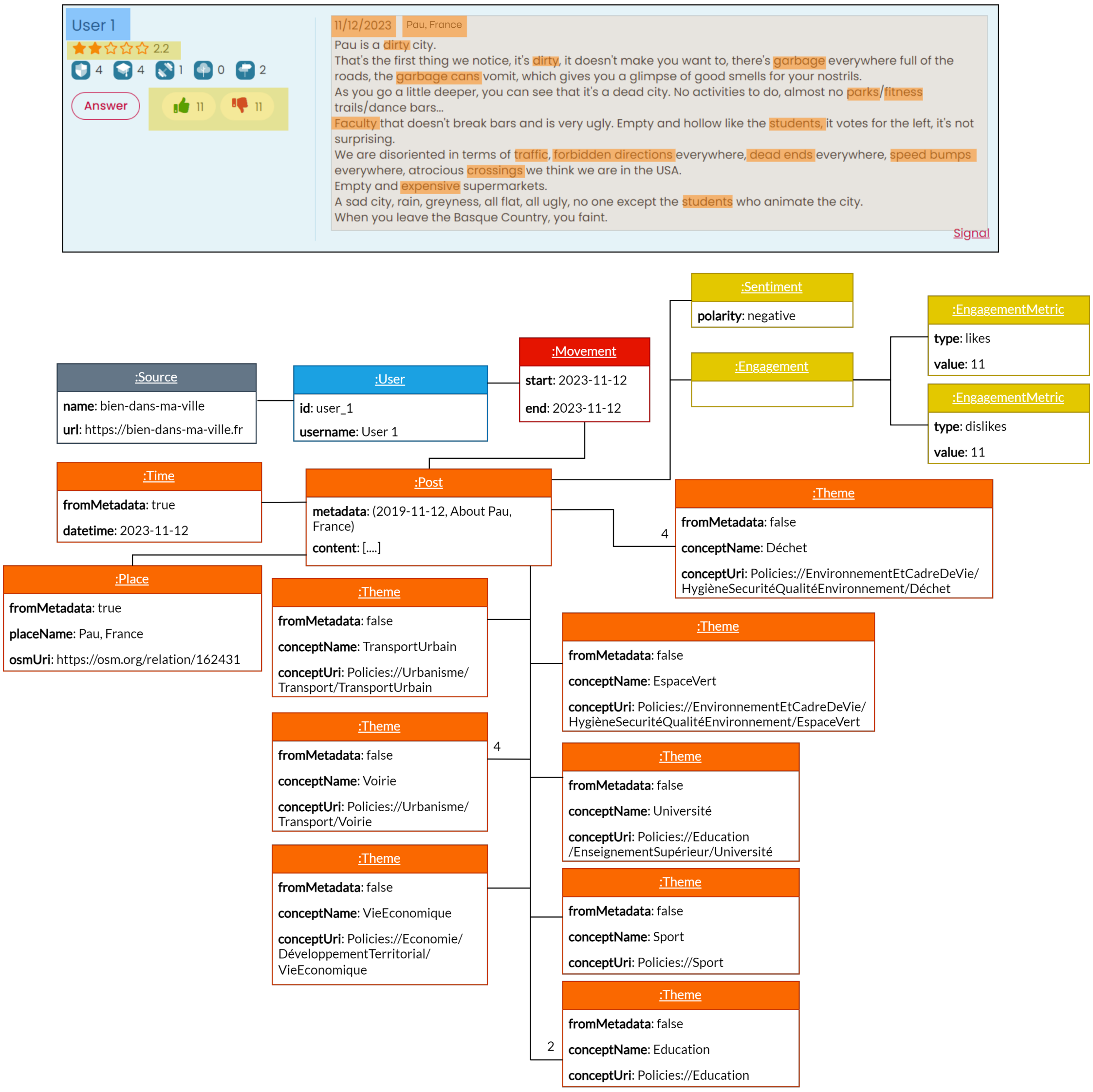
2. For *places*, since all reviews are geotagged, we can reliably determine the location discussed in the review. Our manual analysis indicates that reviews infrequently reference other places; the only exceptions noted were comparisons between municipalities or mentions of specific suburbs or points of interest. For this experiment, we chose to focus solely on the main municipality under review, as it is the primary subject of interest.

3. For *thematic concepts*, we leveraged the rule-based approach previously referenced in Chapter 3. As demonstrated in Figure 3.8, this approach achieves an F1-score comparable to machine learning techniques for this task in the tourism domain. We hypothesize similar results in the domain of local public policies.

The use of these alternative knowledge extraction techniques underscores the flexibility of the *Transform* phase, demonstrating that it can be effectively accomplished through a variety of methods, not limited to the deep learning models we experimented with in the tourism domain.

For the remaining dimensions, such as user, time, and engagement, we extract these variables from the review metadata following the same methodology applied in the tourism experiment.

Using the extracted data, we can instantiate the APs Trajectory Model. Figure 6.4 shows an object diagram of the model instantiated with a review about the municipality of *Pau* (note that it is the same review as in Figure 6.3, for space reasons, we do not display all class attributes). We are therefore able to validate the hypothesis that our data model, based on a formal redefinition of *Proxemics*, can accommodate another domain of interest and data source.



User 1

★★☆☆☆ 2.2

4 4 1 0 2

Answer 11 11

11/12/2023 Pau, France

Pau is a dirty city.
That's the first thing we notice, it's dirty, it doesn't make you want to, there's garbage everywhere full of the roads, the garbage cans vomit, which gives you a glimpse of good smells for your nostrils.
As you go a little deeper, you can see that it's a dead city. No activities to do, almost no parks/fitness trails/dance bars...
Faculty that doesn't break bars and is very ugly. Empty and hollow like the students, it votes for the left, it's not surprising.
We are disoriented in terms of traffic, forbidden directions everywhere, dead ends everywhere, speed bumps everywhere, atrocious crossings we think we are in the USA.
Empty and expensive supermarkets.
A sad city, rain, greyness, all flat, all ugly, no one except the students who animate the city.
When you leave the Basque Country, you faint.

Signal

:Sentiment
polarity: negative
:EngagementMetric
type: likes
value: 11
:Engagement
:EngagementMetric
type: dislikes
value: 11
:Movement
start: 2023-11-12
end: 2023-11-12
:Source
name: bien-dans-ma-ville
url: https://bien-dans-ma-ville.fr
:User
id: user_1
username: User 1
:Time
fromMetadata: true
datetime: 2023-11-12
:Post
metadata: (2019-11-12, About Pau, France)
content: [....]
:Theme
fromMetadata: false
conceptName: Déchet
conceptUri: Policies://EnvironnementEtCadreDeVie/HygièneSecuritéQualitéEnvironnement/Déchet
4
:Place
fromMetadata: true
placeName: Pau, France
osmUri: https://osm.org/relation/162431
:Theme
fromMetadata: false
conceptName: TransportUrbain
conceptUri: Policies://Urbanisme/Transport/TransportUrbain
:Theme
fromMetadata: false
conceptName: EspaceVert
conceptUri: Policies://EnvironnementEtCadreDeVie/HygièneSecuritéQualitéEnvironnement/EspaceVert
:Theme
fromMetadata: false
conceptName: Voirie
conceptUri: Policies://Urbanisme/Transport/Voirie
4
:Theme
fromMetadata: false
conceptName: Université
conceptUri: Policies://Education/EnseignementSupérieur/Université
:Theme
fromMetadata: false
conceptName: VieEconomique
conceptUri: Policies://Economie/DéveloppementTerritorial/VieEconomique
:Theme
fromMetadata: false
conceptName: Sport
conceptUri: Policies://Sport
:Theme
fromMetadata: false
conceptName: Education
conceptUri: Policies://Education
2

Figure 6.4: Instantiation of the APs Trajectory Model Using a Municipality Review from *bien-dans-ma-ville.fr*.

We will now move on to calculate indicators on the instantiated model using the *ProxMetrics* toolkit and associated formula.

6.5 Calculating Indicators on Local Public Policies with *ProxMetrics*

We are currently in the *Analyze* phase of the APs Framework. We have collected a dataset of municipality reviews about local public policies and instantiated our framework's data model with it. We now want to ascertain whether our proxemic similarity toolkit (*ProxMetrics*, Contribution 3.3) can calculate meaningful indicators on the domain of local public policies based on municipality reviews. The stakeholders we will rely on for our experimentation are management researchers in local public policies from the OPTIMA research chair.

We will first introduce requirements from these domain stakeholders and propose a proxemic environment to model them (Subsection 6.5.1), then we will propose three case studies on selected requirements (Subsection 6.5.2, Subsection 6.5.3, and Subsection 6.5.4).

6.5.1 Modeling Proxemic Indicators for Local Public Policies

The initial step involved compiling a list of requirements from management researchers specializing in local public policy (excluding computer scientists) who want to analyze municipality reviews. For this experiment, we will focus on five of their requirements. They are seeking indicators on:

1. *Citizen Satisfaction with Key Local Public Policies*: Evaluate how satisfied citizens are with crucial local public policies in their municipalities. Key areas of interest include safety, health, social services, and environmental policies.

2. *Demographic Similarities Among Municipalities*: Identify municipalities with comparable demographic profiles (e.g., similar age, social professional categories, etc.).

3. *Analysis of Citizen Migration Patterns*: Examine the origins and destinations of migrating citizens, along with the reasons behind their relocation decisions.

4. *Association of Local Public Policies in Citizen Discourse*: Investigate the correlations between different local public policies as they are commonly referenced together in discussions by municipality residents.

5. *Network of Citizens with Shared Policy Concerns*: Facilitate connections among citizens who have similar concerns about local public policies, promoting community engagement and collaborative approaches to problem-solving.

Table 6.2 shows a list of these requirements and how we can model them using our proxemic entities. We have kept the same table format as in Table 4.5 to show the adaptability of the toolkit in modeling requirements in various domains. As a reminder, proxemics requires the selection of a central identity (the reference) and target entities scattered around it. These entities can be dynamic (e.g., users, groups) or static (informational entities found in posts such as places, themes, or periods).

For *Requirement 1* (Table 6.2, 1), we consider a reference municipality and examine local public policies frequently associated with it. The analysis leverages the *Location* dimension, adjusted by sentiment in the *Orientation* dimension. Policies perceived positively and frequently co-occurring with the reference municipality are considered similar, while those perceived negatively or simply missing in the reference municipality are considered dissimilar.

	Proxemic Environment		Dimensions				
Requirement	**Reference (E_{ref})**	**Targets (τ)**	**D**	**I**	**L**	**M**	**O**
1	📍 Municipality	📖 Local Public Policies			•		•
2	📍 Municipality	📍 Municipalities		•	•		•
3	📍 Municipality	📍 Municipalities	•			•	
4	📖 Local Public Policy	📖 Local Public Policies			•		
5	👤 Citizen	👤 Citizens	•	•	•	•	•

Table 6.2: Example of End Users Requirements for Analysis of Local Public Policies

For *Requirement 2* (Table 6.2, 2), the similarity between a reference municipality and target municipalities is primarily based on the similarity of profiles of citizens mentioning them (*Identity* dimension). The *Location* and *Orientation* dimensions are also considered to enrich the proxemic similarity measure, as co-occurrences in reviews, especially positive ones, may hint at higher similarity between municipalities.

For *Requirement 3* (Table 6.2, 3), we use the same proxemic environment but with different dimensions used. Here, we identify municipalities to which citizens tend to migrate from a given reference one, using the *Movement* dimension to identify frequently sequenced municipalities in citizens' trajectories. The *Distance* dimension is also considered, as closer municipalities should weigh more in the analysis. This requirement shows that by modulating proxemic dimensions in different ways, we can model various requirements with the same proxemic environment.

Requirement 4 (Table 6.2, 4) involves using the *Location* dimension to identify public policies frequently mentioned together in reviews. This simpler analysis focuses on the co-occurrence of policy mentions within the same context, providing insights into how different policies are associated with citizen discourse.

Finally, for *Requirement 5* (Table 6.2, 5), we aim to build a system that connects citizens with similar concerns about local public policies. This involves using all the DILMO dimensions. We consider physically close citizens (*Distance* dimension) and connect citizens with similar demographic features (*Identity* dimension). Citizens who discuss the same themes, municipalities, and periods are considered more similar (*Location* dimension), as well as those who speak positively about the same issues (*Orientation* dimension). Additionally, people who share similar migration patterns, such as moving from one municipality to another, are also connected (*Movement* dimension). The weighting of each dimension depends on the priorities set by the domain stakeholders.

We propose to go into detail on three of these requirements, namely *Citizen Satisfaction with Local Public Policies* (*Requirement 1*), *Analysis of Citizen Migration Patterns* (*Requirement 3*), and *Association of Local Public Policies in Citizen Discourse* (*Requirement 4*). We chose these three requirements because they do not require the *Identity* dimension. The municipality review platform we gathered the dataset from (e.g., *bien-dans-ma-ville.fr*) does not provide any profile information beyond the citizens' names and IDs. Therefore, the identification of citizens' personal features would need to be based on their reviews' content, potentially using NLP techniques to categorize citizens into groups, such as social-professional (e.g., students, working professionals, retirees) or familial (e.g., singles, couples, parents, etc.) demographics. The design of such techniques requires significant investigation and falls outside the scope of this experiment, which aims to demonstrate the *ProxMetrics* toolkit's genericity. Let's move to the first case study. It is modeled using proxemic similarity *Pattern 4* (static entity to static entities, refer to Figure 4.7), which we did not conduct a detailed case study on in Section 4.7.

6.5.2 Case Study 1: Citizen Satisfaction with Local Public Policies

In Figure 6.5, the *Citizen Satisfaction with Local Public Policies* indicator is expressed using a proxemic environment falling into *Pattern 4*. The objective of this indicator is to inform local public policy stakeholders (e.g., town councils) about areas of public policy that require improvement, enabling them to take appropriate action.

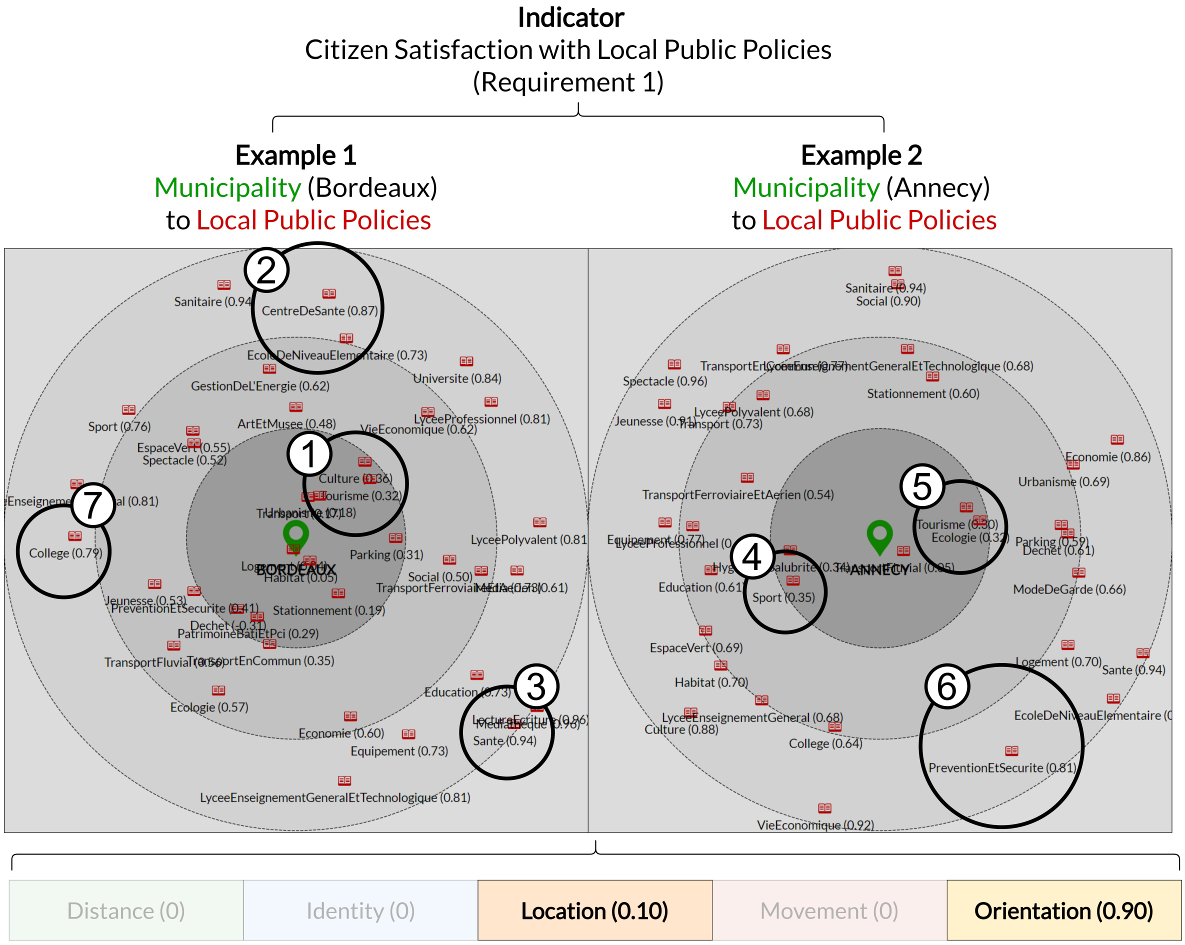


Figure 6.5: Visualization of the Indicator "*Citizen Satisfaction with Local Public Policies*" in the Proxemic Reticle

A municipality of interest is set as a reference (here, we have two examples: *Bordeaux* and *Annecy*), and themes corresponding to public policies are positioned relative to it. By default, we set proxemic dimensions L to $\frac{1}{10} = 0.1$ and O to $\frac{9}{10} = 0.9$ weighting. This is because this requirement is focused on satisfaction, and therefore positive local public policies should weigh more in the similarity. However, this default weighting can be dynamically changed by end-users if they wish to give more weight to local public policy mentions or have the latter weighted by sentiment.

Here is how the proxemic similarity is calculated in this case:

- The *Location* (L) dimension is based on how often the reference municipality co-occurs with local public policies in citizens' reviews, which may indicate a specific affinity of the public policy for the municipality due to their frequent association.
- The *Orientation* (O) dimension is considered and weights municipality mentions by sentiment values, meaning positive co-occurrences will impact the similarity results (e.g., similar results will correspond to local public policies which are often positively mentioned).

Results in Figure 6.5 show that:

- In the case of *Bordeaux*, the municipality seems to be liked for its tourism-related and culture-related policies as they are positioned close to the reference (see Figure 6.5, ①). On the other hand, health-related policy concepts like *CentreDeSante* and *Sante* are quite far due to the negative sentiments expressed about them (refer to Figure 6.5, ② and ③).
- In the case of *Annecy*, tourism and sport-related policy concepts are quite close to the reference (see Figure 6.5, ④), indicating that constituents of this city seem to be satisfied with them. However, *PreventionEtSecurite* is quite far, suggesting that the municipality is seen as lacking in this aspect (see Figure 6.5, ⑤ and ⑥).

Let's illustrate this example using the pairing $E_{ref} = Bordeaux$ and $E_{target} = College$ (refer to Figure 6.5, ⑦). We refer back to Table 4.4 to determine the formula for this proxemic environment pattern, specifically *Pattern 4*. We calculate the proxemic similarity between the entities as follows, using the $L_{co-occurrences}$ and $O_{co-occurrences}$ formula corresponding to the *Location* and *Orientation* (LO) dimensions considered in this example. We obtain a proxemic similarity of 0.79 by aggregating the L and O dimensions. It appears that the municipality of *Bordeaux* does not satisfy citizens in regards to middle school level education (e.g., *Collège*).

$$\begin{gathered} P_s(Bordeaux, College) = \\ \frac{1}{10} \times \underbrace{L_{co-occurences}(Bordeaux, College)}_{0.87} + \frac{9}{10} \times \underbrace{O_{co-occurences}(Bordeaux, College)}_{0.77} \\ \approx 0.79 \end{gathered} \tag{6.1}$$

It is important to acknowledge that we made some adjustments to the $L_{co-occurrences}$ formula used here compared to the initial formula presented in Subsection 4.6.6. More precisely, we changed the λ parameter in the $w_{freshness}$ formula, which affects the weight of reviews based on how fresh (recent) they are. Initially, it was set to -0.01, which worked well in the tourism domain (the dataset ranged over 3 months), but for a dataset like this one, which ranges over 8 years, the value was not appropriate as it made reviews issued more than a year and a half ago weight zero. We use a value of λ of -0.0003 to produce a more balanced effect, as shown in Figure 6.6.

In Figure 6.5, we have defined three proxemic zones corresponding to local public policies associated with the best satisfaction, average satisfaction, and worst satisfaction. This choice was arbitrary; other zones could be defined depending on the results.

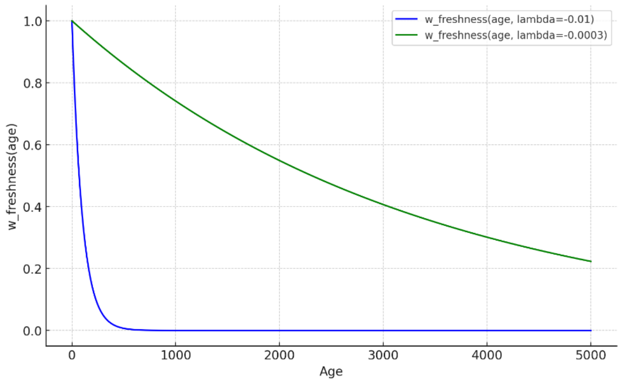


Figure 6.6: Plot Showing the Values of $w_{freshness}$ Depending on the Freshness of Reviews

6.5.3 Case Study 2: Analysis of Citizen Migration Patterns

Now, we present the indicator corresponding to the *Analysis of Citizen Migration Patterns*.

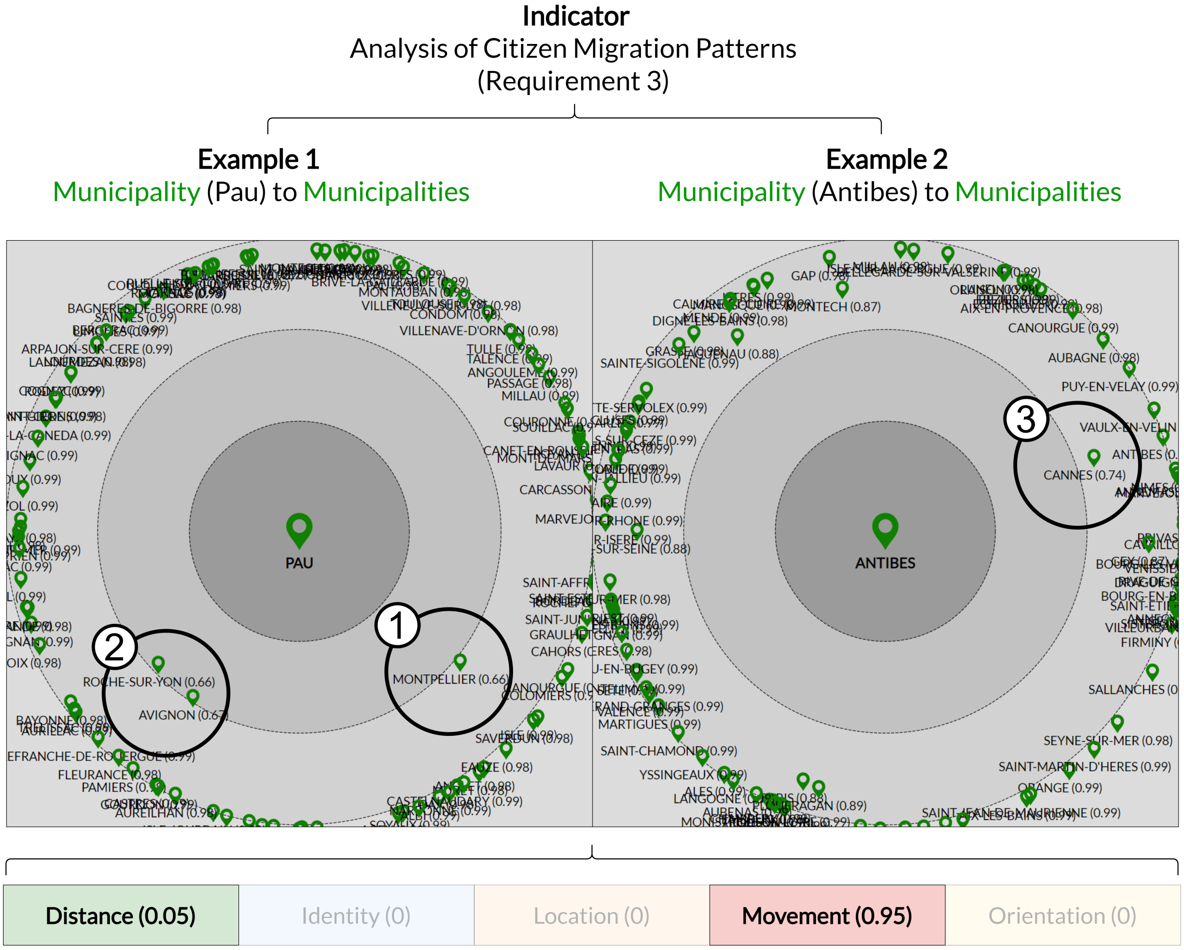


Figure 6.7: Visualization of the Indicator *"Analysis of Citizen Migration Patterns"* in the Proxemic Reticle

This requirement involves investigating where people tend to migrate (i.e., move to) from a given reference municipality. This information helps inform local stakeholders about which municipalities are attracting their former constituents, potentially guiding local public policies to better meet their requirements.

In Figure 6.7, we provide two examples using the municipalities of *Pau* and *Antibes* as reference entities. These municipalities were chosen because they are situated in vastly different areas of *France*. The surrounding municipalities indicate where residents of *Pau* and *Antibes* tend to migrate. The closer a municipality is to the reference municipality, the more people from that municipality tend to move there. It is important to note that the *ProxMetrics* toolkit alone cannot investigate the reasons and motivations behind these migrations (e.g., why people moved). For such analyses, we will need additional tools like the *TextBI* platform, which will be discussed later.

We use two proxemic dimensions, *Distance* (D) and *Movement* (M). The M dimension is weighted at $\frac{1}{20} = 0.05$, while the D dimension is weighted at $\frac{19}{20} = 0.95$. This is because, although our primary interest is the sequencing of places in citizens' spatial trajectories, we also consider that places in close physical proximity are more likely to attract movers and should be slightly boosted in significance. In this case study, we hypothesized that when a user writes multiple reviews about different municipalities, it indicates they have relocated from one to another. This conclusion was drawn from our manual analysis of municipality reviews. While some users may write reviews about municipalities they have never visited (e.g., based on news reports, or short visits), this should be a minority and is unlikely to significantly bias the overall, aggregated results.

Results in Figure 6.7 show that:

- In the case of *Pau*, inhabitants tend to relocate to the municipalities of *Montpellier* (see Figure 6.7, ①) and *Roche-sur-Yon* as well as *Avignon* (see Figure 6.7, ②). As a reminder, for *Pattern 4* the movement similarity is based on a conditional probability (e.g., the probability that an inhabitant of *Pau* moves to *Montpellier* or *Roche-sur-Yon* is proportionally higher compared to other municipalities).
- In the case of *Antibes*, the city of *Cannes* clearly stands out, indicating that most inhabitants of *Antibes* tend to move there (see Figure 6.7, ③).

We will illustrate this example using the pairing $E_{\text{ref}} = \text{Antibes}$ and $E_{\text{target}} = \text{Cannes}$. We refer back to Table 4.4 to determine the formula for this proxemic environment pattern, specifically *Pattern 4*. We calculate the proxemic similarity between the entities as follows, using the D_{physical} and $M_{\text{sequencing}}$ formulas corresponding to the *Distance* and *Movement* (DM) dimensions considered in this example. We get a result of 0.74 by aggregating the D and M dimensions. We notice that this value is significantly higher than for the other municipalities' pairings (refer to Figure 6.7, *right side*), indicating that *Cannes* is a privileged moving place for constituents of *Antibes*.

$$\begin{gathered} P_s(Antibes, Cannes) = \\ \frac{1}{20} \times \underbrace{D_{physical}(Antibes, Cannes)}_{0.99} + \frac{19}{20} \times \underbrace{M_{sequencing}(Antibes, Cannes)}_{0.72} \\ \approx 0.74 \end{gathered} \tag{6.2}$$

6.5.4 Case Study 3: Association of Local Public Policies in Citizen Discourse

The indicator shown in Figure 6.8 corresponds to the requirement *Association of Local Public Policies in Citizen Discourse* (*Requirement 4*). In this study, we select a given reference local public policy and observe other policies often mentioned alongside it. Specifically, we have chosen *Logement* (related to housing) and *PreventionAndSecurite* (related to delinquency and security) as references policies (see Figure 6.8, *Example 1* and *Example 2*).

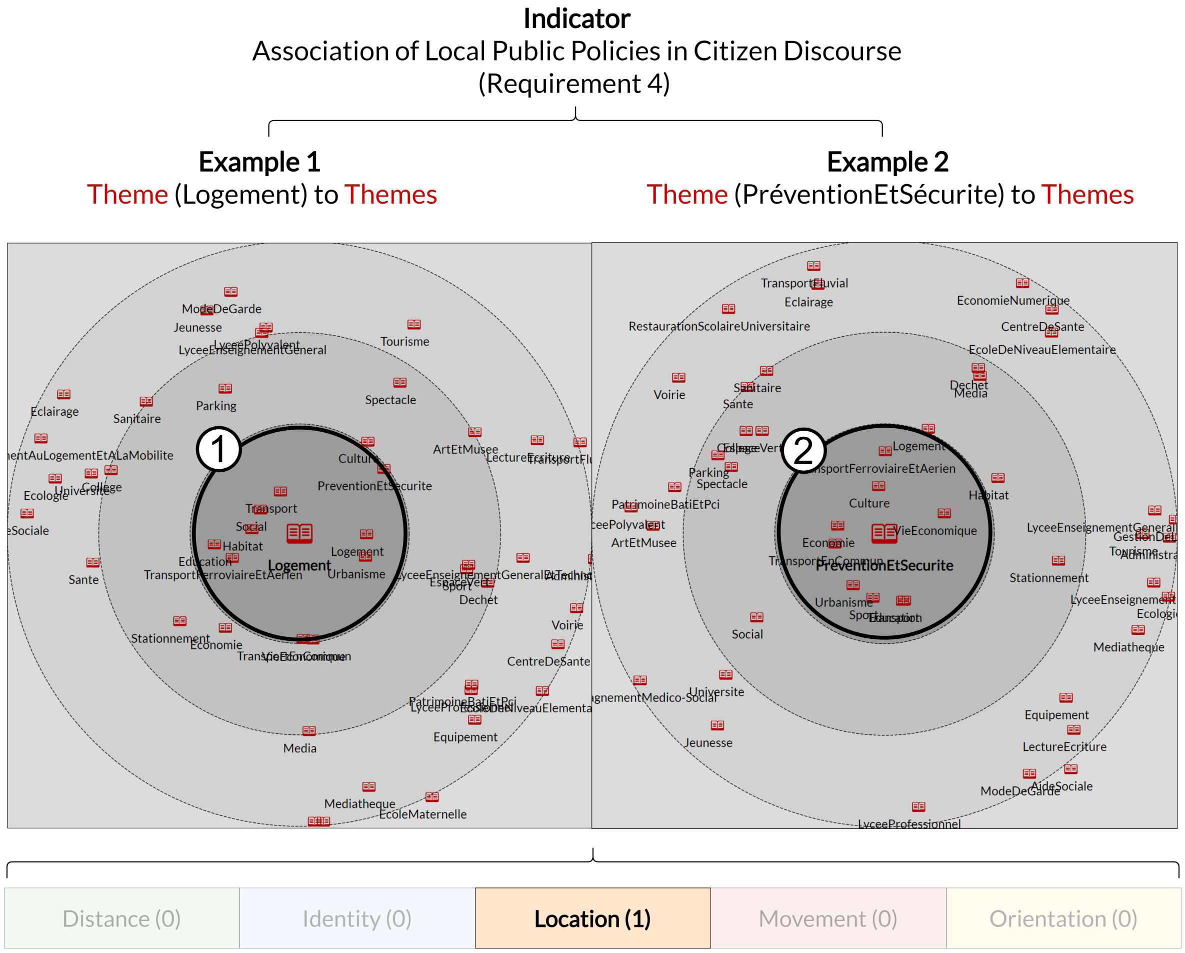


Figure 6.8: Visualization of the Indicator "*Association of Local Public Policies in Citizen Discourse*" in the Proxemic Reticle

This requirement is simpler because it involves a single dimension: the *Location* (L) dimension, which measures the strength of co-occurrences between themes. We will not detail the calculation again, as we did in the two previous case studies, because the calculation here involves only one dimension.

Results in Figure 6.8 show that the housing concept tends to be associated with *Transport*, *Social*, and *Urbanism* (see Figure 6.8, ①), indicating that people often mention commuting-related aspects or social grants for renting apartments. For the *PreventionAndSecurite* policy (see Figure 6.8, ②), it appears that *Transport* is again heavily correlated, along with *TransportFerroviaireEtAérien* and *TransportEnCommun*, potentially indicating concerns about security in public transport.

6.5.5 Discussion

These case studies on the *ProxMetrics* toolkit demonstrate its adaptability to another domain of application: local public policies with different stakeholders' requirements and using another type of data source. This experiment stresses the importance of our proxemic similarity formula parameters. As seen in Subsection 6.5.2, parameters that were suitable for the domain of tourism may need to take different values in different domains. For now, we have set these parameters manually, but it may be interesting to design algorithms that can infer them automatically based on the dataset's nature.

Additionally, an important limitation is the length of municipality reviews. These reviews are very long compared to regular social media posts, which means many different thematic concepts are mentioned, and all of them are aggregated as they are part of the same review, which may cause severe inaccuracies. For example, if a citizen writes a generally positive review but criticises a particular public policy, this policy will be detected as positive when it should not be.

In regular social media posts like on X/Twitter, this is less of an issue because they tend to be very short, and therefore all of the text is usually linked to the same sentiment polarity. In the case of municipality reviews this is the opposite, citizens often speak positively about some local public policies and negatively about others. Therefore, to get finer analyses, several actions are possible:

- Using advanced NLP techniques for Aspect-Based Sentiment Analysis (Barnes et al., 2022; Nazir et al., 2020; Liu et al., 2020) to link specific aspects of the reviews with particular sentiments.
- Using the individual star ratings (see Figure 6.3). These are sub-ratings citizens give to broad families of local public policies (e.g., security, education, sport, etc.). We could associate these ratings with particular branches of our semantic resource (thesaurus of local public policies).

Currently, the APs Trajectory Model does not support deconstructing a post into several aspects associated with different sentiments, which is a significant limitation when working with longer texts. To enable our analyses to be accurate in domains associated with longer texts, such as local public policies, substantial modifications are necessary. One possible approach could involve a system to divide texts into subtexts, each associated with particular places, periods, themes, or sentiments. For now, let's load the new dataset into the *TextBI* platform.

6.6 The *TextBI* Dashboard Applied to Local Public Policies

We will now demonstrate the adaptability of the *TextBI* platform with this new domain and data source. In contrast to the previous chapter (refer to Chapter 5), where the entire dashboard was presented, here we will focus on showcasing selected visualizations (see Subsection 6.6.1) and evaluate with another type of user, management researchers (see Subsection 6.6.2), to determine whether their assessment is similar to that of tourism stakeholders.

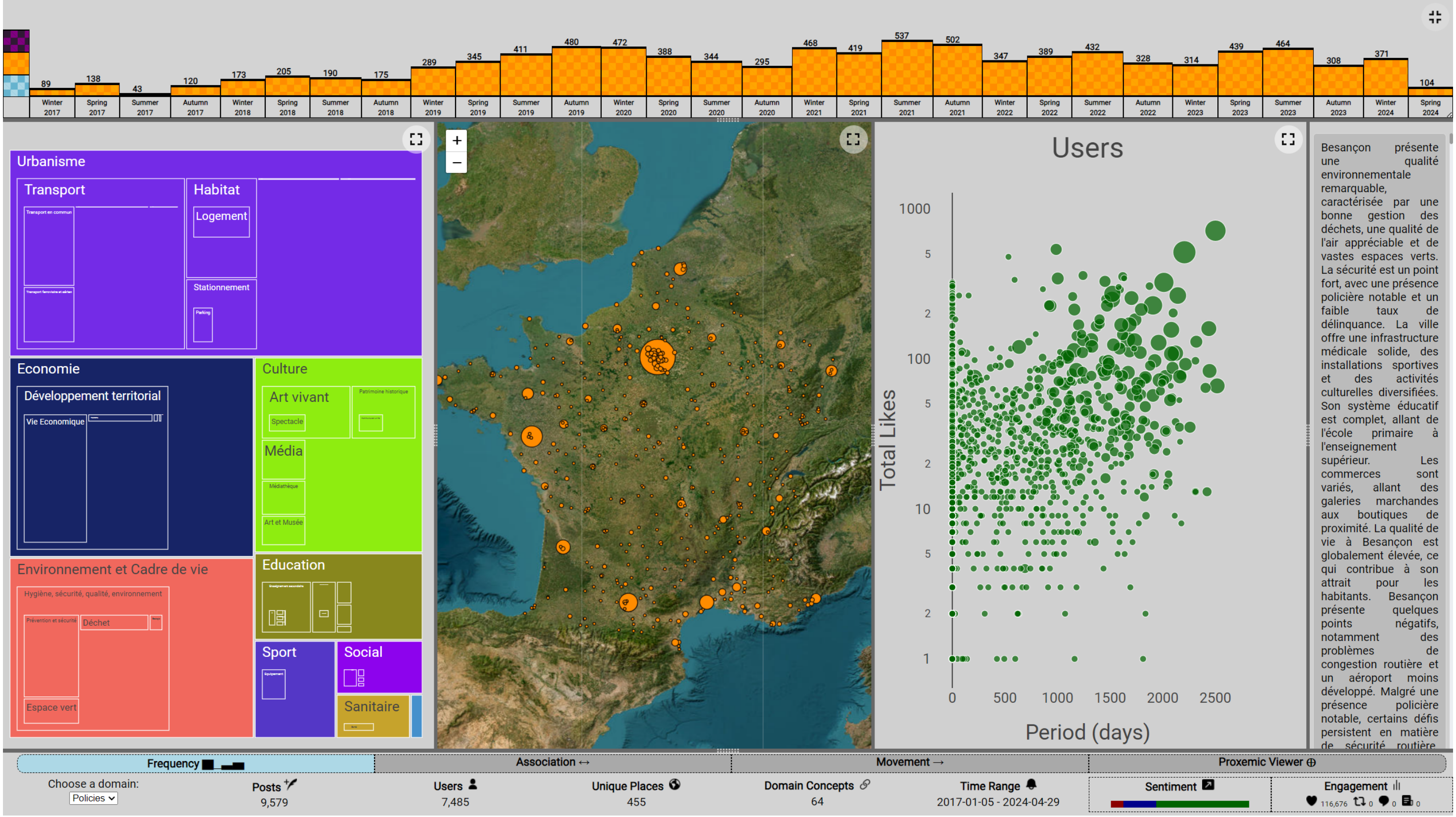
6.6.1 Experimentation: *TextBI* to Visualize Analyses on Local Public Policies

Firstly, Figure 6.9 shows a screenshot of the *Frequency* view loaded with the local public policies dataset. We observe that the most frequently mentioned local public policies are related to *Urbanism* (depicted in purple in the thematic map) and *Economy* (depicted in blue in the thematic map).

The *Spatial Map* is not a choropleth one but a bubble map, as the dashboard detected the high number of municipalities and adapted the visualization accordingly. Most of the reviews originate from *Paris*, *Nantes*, and the southeastern Mediterranean coast of France.

The *User Map* uses only one type of color and symbol because all reviews are in French, unlike touristic tweets which are multilingual. Here, we display the activity time of citizens on the x-axis, which is longer than that of visitors in the tourism domain, and the total number of likes received by each user on the y-axis, this allows us to identify influential users.

Finally, the *Timeline* is segmented by season. Local public policy stakeholders informed us that they are mainly interested in this segmentation when assessing the temporal dimension as it allows a balanced temporal distribution of reviews (e.g., feelings about a municipality in summer, winter, etc.). Unlike for tourism, we do not have morning, afternoon, and evening segmentation because the municipality review platform we used does not return this information, only the date at the day level.



Besançon présente une qualité environnementale remarquable, caractérisée par une bonne gestion des déchets, une qualité de l'air appréciable et de vastes espaces verts. La sécurité est un point fort, avec une présence policière notable et un faible taux de délinquance. La ville offre une infrastructure médicale solide, des installations sportives et des activités culturelles diversifiées. Son système éducatif est complet, allant de l'école primaire à l'enseignement supérieur. Les commerces sont variés, allant des galeries marchandes aux boutiques de proximité. La qualité de vie à Besançon est globalement élevée, ce qui contribue à son attrait pour les habitants. Besançon présente quelques points négatifs, notamment des problèmes de congestion routière et un aéroport moins développé. Malgré une présence policière notable, certains défis persistent en matière de sécurité routière

Figure 6.9: The Frequency View Applied to the Local Public Policies Dataset

At the top of Figure 6.10, we can observe in the timeline that the sentiment has degraded significantly in recent years (the orange color denotes a mixed sentiment). We, therefore, decided to compare the *Thematic* and *Spatial* maps over two time ranges (see Figure 6.10), namely 2017 to 2019 and 2022 to 2023, each spanning three years. As we can see in the *Spatial Map*, the satisfaction of citizens has dropped severely in almost all municipalities of *France*, with prime examples being the municipalities of *Nantes* and *Avignon*.

On the *Thematic Map* side, we can see that most local public policy branches went from positive to mixed, especially *Urbanism* and *Living Environment*. Some, like *Culture*, have stayed strongly positive. The distribution of policies is rather similar, but it appears that *Living Environment* has grown significantly, highlighting a lot of concerns from citizens regarding these policies.

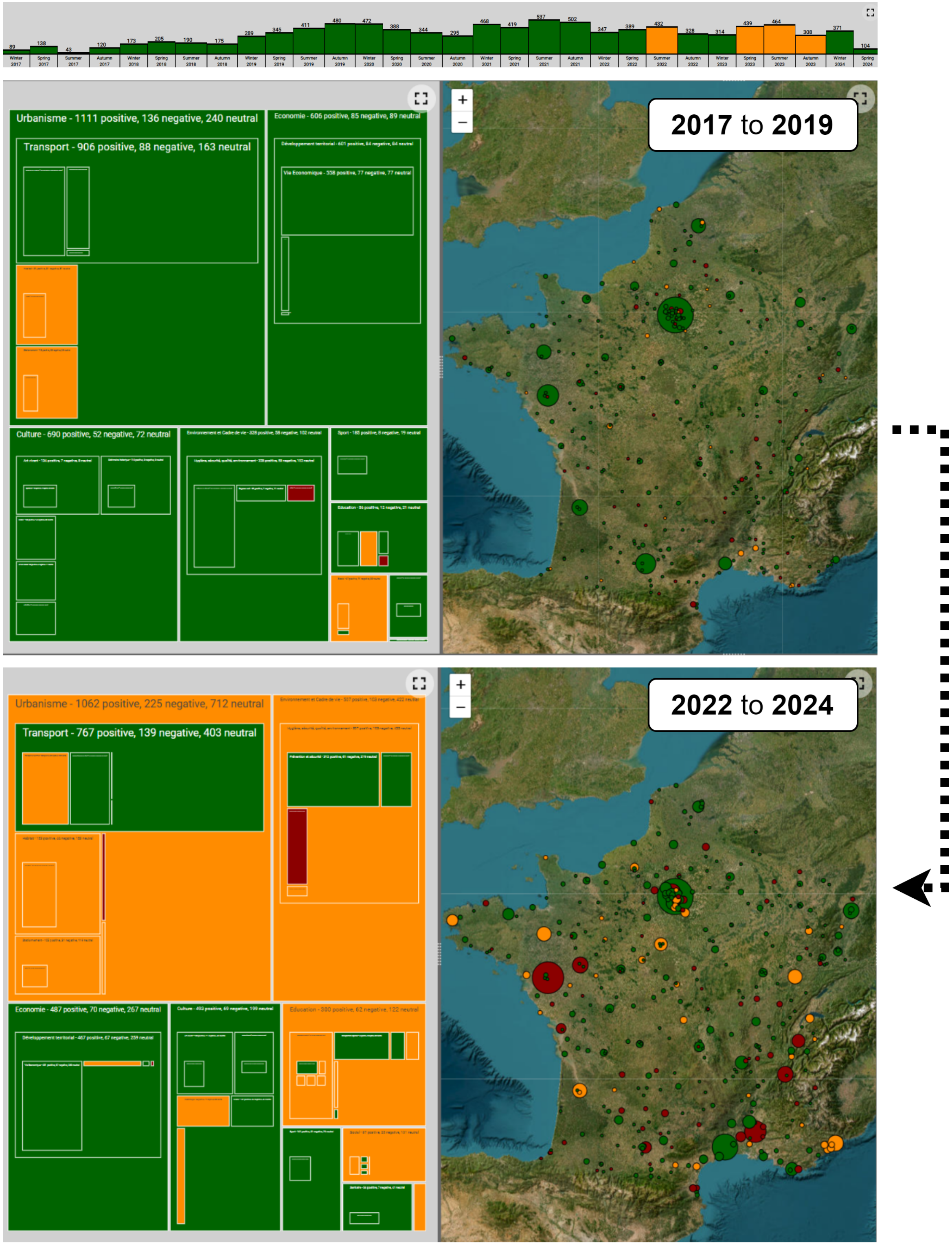

89
Winter 2017
138
Spring 2017
43
Summer 2017
120
Autumn 2017
173
Winter 2018
205
Spring 2018
190
Summer 2018
175
Autumn 2018
289
Winter 2019
345
Spring 2019
411
Summer 2019
480
Autumn 2019
472
Winter 2020
388
Spring 2020
344
Summer 2020
295
Autumn 2020
468
Winter 2021
419
Spring 2021
537
Summer 2021
502
Autumn 2021
347
Winter 2022
389
Spring 2022
432
Summer 2022
328
Autumn 2022
314
Winter 2023
439
Spring 2023
464
Summer 2023
308
Autumn 2023
371
Winter 2024
104
Spring 2024
Urbanisme - 1111 positive, 136 negative, 240 neutral
Transport - 906 positive, 88 negative, 163 neutral
Economie - 606 positive, 85 negative, 89 neutral
Développement territorial - 601 positive, 84 negative, 84 neutral
Vie Economique - 358 positive, 77 negative, 77 neutral
Culture - 690 positive, 52 negative, 72 neutral
Environnement et Cadre de vie - 328 positive, 58 negative, 102 neutral
Hygiène, sécurité, qualité, environnement - 328 positive, 58 negative, 102 neutral
Sport - 185 positive, 8 negative, 19 neutral
Education - 86 positive, 12 negative, 21 neutral
+
−
2017 to 2019
Urbanisme - 1062 positive, 225 negative, 712 neutral
Transport - 767 positive, 139 negative, 403 neutral
Environnement et Cadre de vie - 507 positive, 103 negative, 422 neutral
Hygiène, sécurité, qualité, environnement - 507 positive, 103 negative, 422 neutral
Prévention et sécurité - 212 positive, 61 negative, 219 neutral
Economie - 487 positive, 70 negative, 267 neutral
Développement territorial - 467 positive, 67 negative, 259 neutral
Culture - 493 positive, 69 negative, 199 neutral
Education - 300 positive, 62 negative, 122 neutral
+
−
2022 to 2024

Figure 6.10: Comparison of Sentiments Expressed About Local Public Policies and Municipalities

Figure 6.11 shows the *Thematic Association Graph* with the *Engagement Overlay* enabled (no filter applied). This graph shows the local public policies that are most often associated (co-occurring) in municipality reviews. The strongest association is shown between *Transport* and *Vie Économique*-related policies. They are also the policies that receive the most engagement on municipality review platforms (e.g., reviews containing them are often liked or disliked).

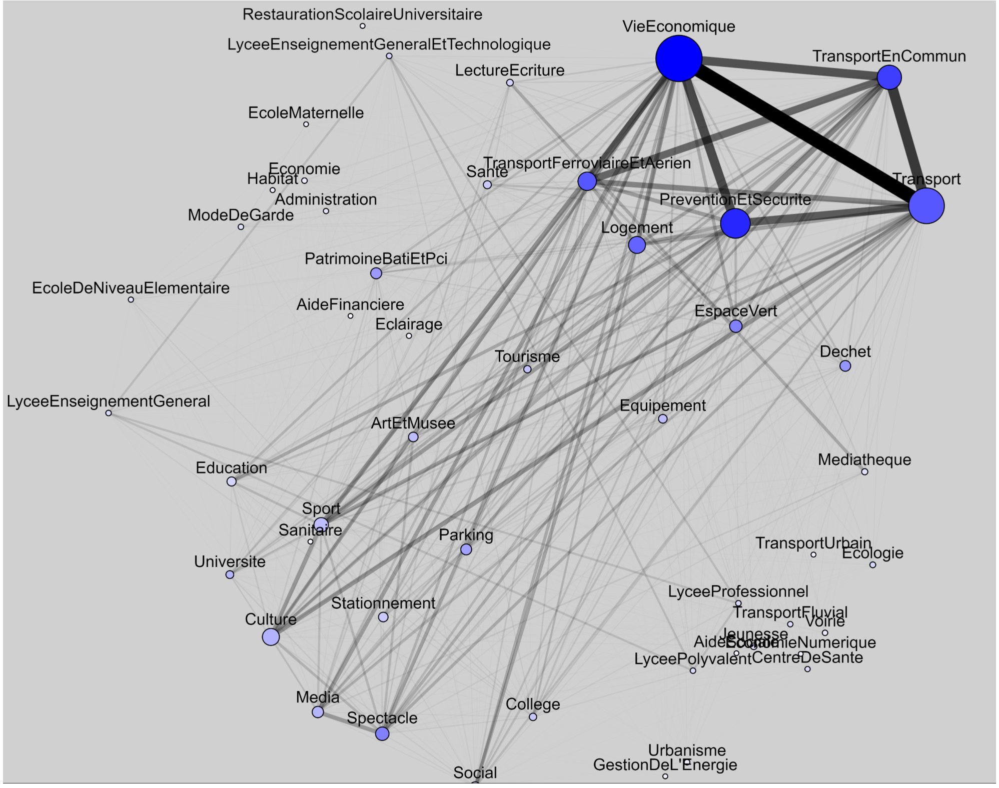


Figure 6.11: Thematic Associations of Local Public Policies with the Engagement Overlay Applied

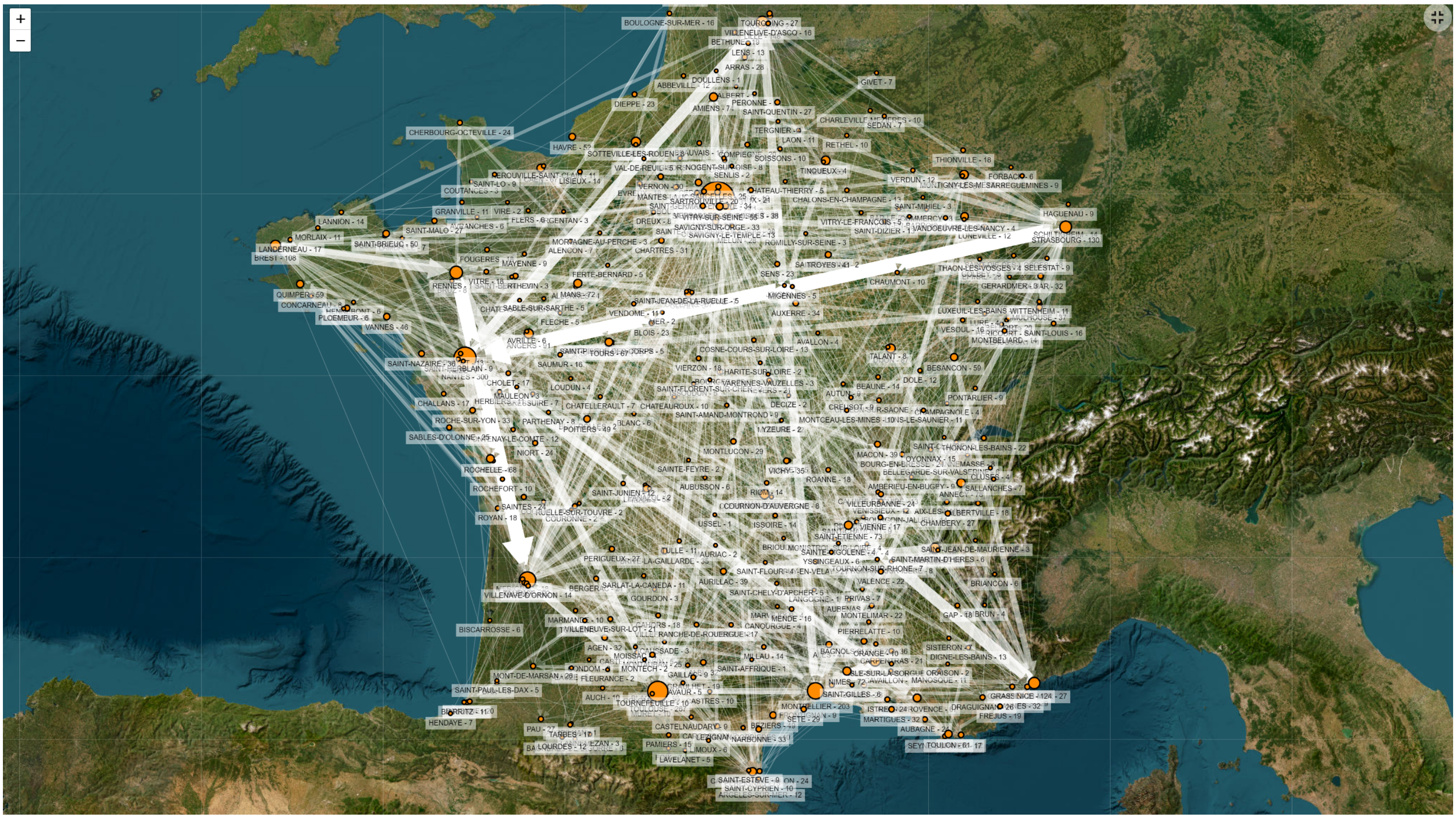


Figure 6.12: Spatial Movement Between Municipalities

Finally, Figure 6.12 shows the spatial movement between municipalities. As we explained previously, we hypothesized that these movements are caused by people relocating from one municipality to another for personal or professional reasons. Major municipalities such as *Nantes*, *Bordeaux*, and *Lyon* show a high concentration of incoming and outgoing citizens, indicated by the thick edges converging at these hubs. The west of *France* around the *Nantes* area appears to experience a lot of relocation.

We will not present the *Proxemics* view again here, as the interface is the same as in the tourism domain, and we have already provided examples of results in the domain of local public policies in the previous sections (see Section 6.5). Instead, let's assess whether the evaluation and opinions of the platform by management researchers in the domain of local public policies are similar to those of the tourism office in the tourism domain.

6.6.2 Qualitative Evaluation of *TextBI* by Stakeholders in Local Public Policies

The qualitative evaluation of *TextBI* with stakeholders in local public policies adheres to the evaluation protocol, criteria, and metrics previously established in the context of the tourism domain, as detailed in Section 5.5. This evaluation is carried out in collaboration with a management researcher specializing in local public policies from the OPTIMA research chair, who serves as the evaluator. The primary objective of it is to determine whether the insights derived from the application of *TextBI* in the tourism domain, with a tourism office director as the end user (presented in Table 5.7), remain applicable and consistent in the context of local public policies with a different end user profile. This is aimed at ascertaining the robustness of the *TextBI* tool across different domains and user profiles.

To this end, a comparative analysis is conducted between the new results obtained in the domain of local public policies and the previous results from the tourism domain. The findings are presented in Table 6.3. For each view, we propose a *Kiviat* diagram comparing the results in the domain of tourism (in blue) and in the domain of local public policies (in green). As a reminder, we are looking for a satisfaction rate of 75% (3.75 out of 5 on the *Likert* scale).

Results in Table 6.3 (see Appendix D for detailed results) show that in the *Frequency View*, the scores for *Feel* are identical, while *Usability*, *Design*, *Features*, and *Performance* show slightly higher scores in tourism, though the difference is negligible. The *Association View* and *Proxemics View* reflect a similar pattern. The *Features* criterion is consistently rated lower in local public policies due to the lack of accuracy previously discussed (caused by longer reviews). Similar observations were made regarding the cluttering of the graph view; as a reminder, the tourism office director preferred a simplified graph showing only highlights. End users require fast, filtering-free access to indicators for reporting purposes without extensive manipulation.

In the *Movement View*, the difference is more pronounced. The *Feel* scores remain equal, but there is a noticeable disparity in *Performance*. We believe this is due to the new dataset being approximately three times larger (≈ 10,000 tweets) and covering a much wider area, thereby complicating the directed graph and extending the loading time to approximately 10 seconds compared to instant loading in the tourism domain (≈ 3,000 tweets). This issue arises solely in the *Movement View*; other views perform satisfactorily. Overall, as we have observed previously, the *Movement View* is seen as inferior to the other views.

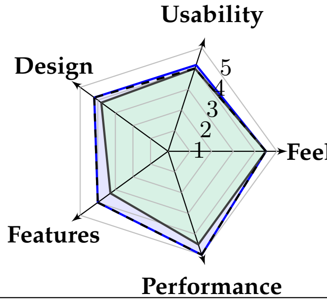
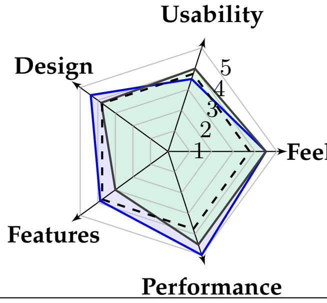
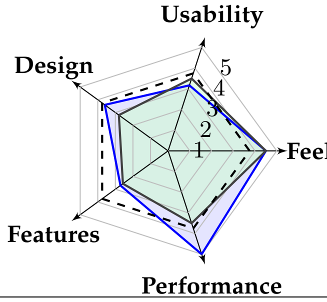
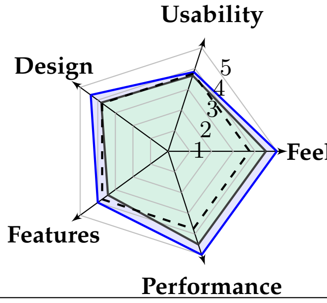
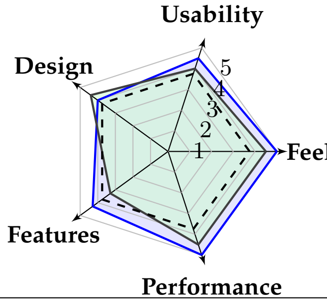
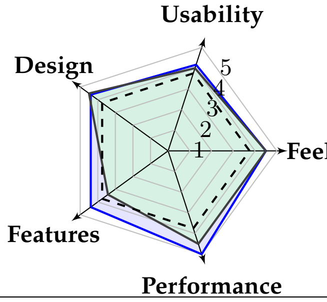
Target	Criteria	Result in Tourism	Result in Public Policies	Comparison
Frequency View	Feel	4.5	4.5	
	Usability	4.17	4	
	Design	4.2	3.8	
	Features	4	3.28	
	Performance	5	4,5	
Association View	Feel	4.5	4.5	
	Usability	3.5	4	
	Design	4.4	3.8	
	Features	3.86	3	
	Performance	5	4,5	
Movement View	Feel	4.5	4.5	
	Usability	3.17	3.5	
	Design	3.6	2.8	
	Features	2.71	2.57	
	Performance	5	3,5	
Proxemics View	Feel	5	4.5	
	Usability	3.83	3.66	
	Design	4.4	3.8	
	Features	4	3.42	
	Performance	5	4.5	
Overlays	Feel	5	4.5	
	Usability	4.5	4	
	Design	4	4.4	
	Features	4.29	3.28	
	Performance	5	4,5	
Global	Feel	4.5	4.5	
	Usability	4.17	4	
	Design	4.4	4.2	
	Features	4.4	3.42	
	Performance	5	4,5	

Table 6.3: Comparison of the Evaluation Results in the Tourism Domain and in the Local Public Policies Domain

It is important to highlight that the overlays, especially the sentiment one, were greatly valued by the management researcher, who was less critical in terms of design because they facilitate the identification of local public policies that require corrective actions and should be prioritized by local authorities (e.g., municipal and regional councils) for improvement.

Having compared the evaluation results of two end users in distinct application domains, we now conclude this experimentation.

6.7 Conclusion

In this chapter, we have demonstrated the applicability of the APs Framework and associated proposals on a dataset of municipality reviews from French municipalities, focusing on the domain of local public policies. The objective was to showcase the genericity of our proposals in both a different application domain and a new data source. This new domain is supported by a newly created semantic resource, a thesaurus of local public policies, developed in collaboration with management researchers due to the lack of existing resources.

The main limitation we encountered was caused by the length of reviews. These are quite long, and our current analysis approach does not allow for analysis as fine-grained as needed, causing potential inaccuracies. For example, currently, all the themes in a long review will be assimilated to the dominant sentiment of the whole review. This was not particularly problematic in regular social media experiments, such as those on X/Twitter, due to the relative shortness of posts which mostly contained single sentiments. However, this issue became pressing with long municipality reviews. Additional work is needed on the model and associated proposals to enhance the accuracy of analysis on social media composed of long texts, such as by integrating support for aspect-based sentiment analysis (Nazir et al., 2020; Liu et al., 2020) to our processing pipeline and model.

Initially, we focused on the *Collect* phase of the framework and applied the iterative collection methodology (Contribution 1) to build a new dataset of municipality reviews about local public policies. The methodology proved effective in constructing a relatively large (9,785 reviews from 7,619 users in 456 municipalities, which is about three times larger than the tourism dataset used in previous chapters) and accurate dataset (accuracy of 0.83 on 50 randomly selected reviews) through various iterations and feedback on the collection filters.

Next, we extracted sentiments, places, and themes from municipality reviews in the *Transform* phase. Unlike the previous experiment in the tourism domain (Contribution 2), we did not use deep learning-based techniques to demonstrate that this step can be achieved with any techniques. We used this dataset to instantiate the APs Trajectory Model (Contribution 3.1 and 3.2), the data model serving as the backbone of our framework. The model proved easily adaptable to this new domain and data source.

In the next phase, *Analyze*, we experimented with the *ProxMetrics* toolkit (Contribution 3.3),, which aims to calculate proxemic similarity indicators to address domain-specific requirements across various domains. The experiment demonstrated that the toolkit can model various indicators corresponding to the requirements of domain stakeholders, in our case, management researchers in local public policies. The necessary adaptations were minor (e.g., changes to the value of some formula parameters), and the core formulas remained relevant.

Then, in the *Valorize* phase, we experimented with the *TextBI* dashboard and demonstrated its genericity to this new data. A qualitative evaluation with stakeholders from this new domain showed that they were mostly in agreement with what was assessed by tourism stakeholders.

We will now conclude this thesis by presenting a retrospective of our proposals and discussing longer-term perspectives to improve this work.

Chapter 7

Conclusion

"What we know is a drop, what we don't know is an ocean."
— Isaac Newton, English Mathematician and Physicist

In this final chapter, we conclude this thesis by first providing an overview of the proposals made in this manuscript (Section 7.1), followed by an exploration of perspectives for extending our work (Section 7.2).

7.1 Thesis Overview

In this section, we offer a comprehensive summary of the key contributions, methodologies, and findings presented in this thesis. For contributions, challenges, and hypothesis numbering, please see Subsection 1.4.3.

7.1.1 Background and Motivations

Since the advent of Web 2.0, User-Generated Content (UGC) has become a predominant data source, particularly UGC coming from social media. Social media are used across a broad and diverse spectrum of application domains. In the domain of *tourism* especially, social media have become very important to complement existing analysis processes. Indeed, social media data presents many advantages including its affordability, diversity, and freshness. However, handling social media data is especially challenging due to its unstructured nature, vastness, and largely multilingual nature.

This thesis is conducted as part of the APs Project, a collaborative French-Spanish project aiming to develop a comprehensive suite of tools for the processing and analysis of social media data, around a semantically defined application domain with a focus on the *Béarn* and *Basque Country* regions, highly touristic cross-border regions. We introduce a generic framework designed for the processing and analysis of social media data: the APs Framework. This framework is generic in two key areas: (1) it is adaptable to various social media (data sources) and (2) applicable across different application domains (although the main application domain in this thesis will be the domain of tourism). The APs Framework encompasses four main phases: *Collect*, *Transform*, *Analyze*, and *Valorize*. The contributions proposed in this thesis are distributed into these phases and cover multiple scientific fields.

The framework architecture has been presented at the *Young Researcher Forum* at INFORSID 2022 (Masson, 2022).

7.1.2 Toward a Generic and Iterative Methodology for Constructing Thematic Datasets from Social Media (Collect)

In the *Collect* phase of the APs Framework, we addressed the challenge of constructing accurate and representative datasets from massive and noisy sources like social media (Challenge 1). This challenge is part of the *Web and Social Media Search* research field. We observed that many research projects rely on data from social media, but most implement ad-hoc techniques for data collection, which are not easily reusable in other application domains or for other purposes. Upon reviewing existing filtering techniques, we noticed that most of the techniques employed were based on three core dimensions: *spatial*, *temporal*, and *thematic*, and two types of data: the *contents* and the *metadata* of the social media posts. This led us to hypothesize that a formalized methodology combining these different criteria could provide a generic and reusable process to build thematic datasets from social media (Hypothesis 1).

The first contribution of this thesis is therefore: a generic and iterative methodology for constructing thematic datasets from social media (Contribution 1). This methodology is based on an iterative and incremental process, incorporating feedback from end-users to construct accurate datasets with neither too much noise nor silence.

This methodology is experimented with via the construction of a thematic dataset on tourism in the *French Basque Coast* using X/Twitter and is evaluated using both quantitative metrics (e.g., comparisons with X/Twitter annotations) and qualitative ones (e.g., evaluation of the accuracy by domain experts). These experiments conducted in the domain of tourism demonstrated the effectiveness of the methodology to build an accurate (e.g., accuracy of 0.74 for geotagged posts, 0.65 for others with only 3 iterations) and a sizeable (e.g., 27,379 posts) dataset.

The collection methodology has been presented at the 2022 edition of the *Web Information System Engineering* (WISE) international conference (Masson et al., 2022).

7.1.3 Optimal Strategies for the Multilingual Analysis of Social Media Content in the Tourism Domain (Transform)

The *Transform* phase of the APs Framework is positioned within the research fields of *Natural Language Processing*, *Information Extraction (IE)*, and *Multilingual Text Mining*. Here, we are faced with two key challenges: transforming unstructured textual data into structured knowledge (Challenge 2) and extracting the latter from multilingual texts by leveraging a minimal number of annotated examples while maintaining competitive performance (Challenge 3). More precisely, we attempted to determine the optimal strategies for multilingual data analysis from social media around three common Information Extraction (IE) tasks in this domain: Sentiment Analysis, Named Entity Recognition (NER) for Locations, and Fine-grained Thematic Concept Extraction.

We reviewed existing techniques, including those based on rules (e.g., lexicon, patterns, grammar, semantics) and deep learning (e.g., fine-tuning, few-shot prompting, existing training corpora) to achieve these tasks.

Firstly, due to the absence of manually annotated datasets with contextual information related to the tourism domain, we hypothesized that a multilingual annotated dataset can establish a solid foundation for processing multilingual social media data in the tourism domain using deep learning techniques (Hypothesis 3). This dataset constitutes the first part of the second contribution (Contribution 2.1), it is based on a subset of posts extracted in Chapter 2 and has been manually annotated at document level with sentiments and at token level with places and fine-grained concepts related to the *Thesaurus on Tourism and Leisure Activities of World Tourism Organization* (World Tourism Organization, 2002). To the best of our knowledge, this is the first dataset of its kind in the tourism domain.

Secondly, we hypothesized that for each specific domain of application, a comparative analysis among the existing NLP techniques and language models would allow for the identification of the most suitable ones for this domain (Hypothesis 2). Therefore, we conducted a comparative analysis on the three information extraction tasks (Contribution 2.2) using several techniques, notably, rule-based, fine-tuning, and few-shot prompting (e.g., prompt-based, Pattern-Exploiting Training (Schick and Schütze, 2021), EntLM (Ma et al., 2022), SetFit (Tunstall et al., 2022)) with various language models (both MLMs like XLM-RoBERTa (Conneau et al., 2019), mBERT (Devlin et al., 2019) and LLMs like Mistral (Jiang et al., 2023), LLaMa 2 (Touvron et al., 2023b), GoLLIE (Sainz et al., 2024), or the GPT series (Brown et al., 2020)) and data sampling techniques. The goal was not only to determine the best strategies but also, for deep learning-based techniques, to determine how many annotations are truly necessary to achieve good results in the domain of tourism, thus avoiding lengthy and costly data annotation efforts.

Extensive experimentation comparing few-shot and fine-tuning techniques with multilingual language models demonstrated that modern few-shot techniques allow us to obtain competitive results for all three tasks with little annotation data: 5 tweets per label (15 in total) for Sentiment Analysis, 30 examples of locations for NER and 1k tweets annotated with fine-grained thematic concepts. We believe that these findings, grounded in a novel dataset, are helpful not only for the development of our application but also for other domain-specific applications, which may require NLP techniques for model enrichment, especially when there is a lack of annotated data or when one wishes to exclude ad-hoc rule-based approaches.

Both the dataset and comparative analysis on NLP techniques have been presented at the 2024 edition of the *Computer Science for Organizations and Information and Decision Systems* (INFORSID) national conference (Masson et al., 2024a)

7.1.4 Redefining *Proxemics* to Model Social Media Entities and their Interactions to Generate Domain-Adaptable Indicators from Social Media (Analyze)

The *Analyze* phase of the APs Framework is positioned within the scientific fields of *Decision Support Systems* and *Data Analytics*. The challenge here was to design meaningful domain-adaptable indicators for actionable insights for domain stakeholders to address requirements across various domains (Challenge 4).

We reviewed existing works on modeling and calculating indicators on social media. From our review, it appears that most existing works are either too focused on a specific domain of application or propose data models and metrics that are too specific, and therefore, not modular enough to accommodate the wide range of requirements from domain stakeholders. Following this, we reviewed the existing uses of the *proxemics* theory (e.g., the science that studies the effect of space and distance on social interactions), which is traditionally used in physical spaces and with physical metrics. We hypothesized that adapting the *proxemics* theory (Hall, 1966) and proxemic dimensions (Greenberg et al., 2011) to social media could provide a generic and versatile way to model interactions and to produce relevant domain-adaptable indicators from social media data (Hypothesis 4).

The contribution here is threefold. Firstly a formal redefinition of this theory and its associated dimensions (Distance, Identity, Location, Movement, and Orientation) for use in digital social media spaces (Contribution 3.1). Based on this redefinition, we also proposed a proxemic trajectory data model, the APs Trajectory Model (Contribution 3.2) along with OCL[1] constraints. This data model allows for modeling social media entities, along with their trajectories and interactions, in a domain-independent manner based on proxemic dimensions. It is designed to be modular and extensible, to accommodate a wide range of use cases and requirements. Unlike existing social media models, it also incorporates the concept of proximity into digital social media spaces. Lastly, based on this redefinition of *proxemics* and data model, we introduced *ProxMetrics*. A toolkit for generating composite, domain-adaptable indicators from social media (Contribution 3.3). It enables the expression of indicators as proxemic similarity metrics between multidimensional social media entities. These indicators are multi-criteria and customizable to be domain-adaptable. They allow for the modulation of the five dimensions (Distance, Identity, Location, Movement, Orientation, DILMO) of *proxemics* to address various domain requirements. The formal redefinition of *proxemics* and associated model were qualitatively evaluated through the modeling of a tweet corpus on the *French Basque Coast* region and around the *tourism* domain. The proxemic trajectory model was instantiated using the most efficient NLP techniques highlighted in the previous chapter. We then collected requirements of local tourism stakeholders (namely, the *Tourism Office of the Basque Country*[2]) to experiment with the toolkit to generate indicators relevant to their requirement. The results of the toolkit were evaluated by human evaluators, and in the majority of the cases, the results given correlated with the assessment of human evaluators (15 cases out of 18).

The proposals of this chapter have been published at the 2023 edition of the *Symposium on Intelligent Data Analysis* (IDA) international conference (Masson et al., 2023b) and in the *Social Network Analysis And Mining* international journal (Masson et al., 2024b).

7.1.5 Interactive Visualization of Multidimensional Analyses from Social Media (Valorize)

The *Valorize* phase of the APs Framework is positioned in the research fields of *Visualization*, *Human-Computer Interactions (HCI)*, and *Interactive Systems and Tools*. Here, the challenge was in presenting indicators and the results of complex social media analyses to non-computer scientists users (namely, stakeholders in various domains) as well as making the findings visually accessible and understandable (Challenge 5).

We reviewed common platforms for this type of decision-support visualization, focusing on four main areas: Domain-Specific Dashboards, Geographic Information Systems (GIS), Business Intelligence (BI), and Linguistic Information Visualizations. We observed that all categories of tools present interesting features in relation to our requirements, but none can address all of them,

[1] *Object Constraint Language*

[2] https://www.en-pays-basque.fr

especially in a way that is suitable for non-computer scientists users. We hypothesized that an interactive dashboard based on four broad dimensions: spatial, temporal, thematic, and personal correlated with enrichment data such as sentiment and engagement, inspired by design elements from existing tools, could provide an accessible and versatile way for non-computer scientist users to analyze social media data and social media-based indicators in various domains of application (Hypothesis 5).

We introduced an interactive platform: *TextBI*[3] (Contribution 4). The platform combines various elements from the aforementioned tools: it incorporates advanced spatial views from GIS, leverages the advanced interactivity and combined filtering capabilities of BI tools, and integrates text-based aspects from Linguistic Information Visualizations into a coherent multidimensional dashboard. Moreover, *TextBI* incorporates the previously calculated proxemic similarity measures and presents them in a user-friendly manner to end-users.

The platform underwent qualitative evaluation through a collaboration with a local tourism office. This evaluation was carried out through a multi-aspect survey where they evaluated various aspect of the platform.

The *TextBI* platform was presented at the 2024 edition of the *European Chapter of the Association for Computational Linguistics: System Demonstrations* (EACL) international conference (Masson et al., 2024c), in the national journal *Mappemonde* (to be published), and in the workshop *Exploring Traces in an All-Digital World: Challenges and Perspectives* at *INFORSID 2023* (Masson et al., 2023a). It also won 1st place at the *Geodata Challenge* of the *National Geonumeric Days*[1] (*GeoDataDays 2023*).

7.1.6 Application of the APs Framework to Another Domain and Data Source

We further demonstrated the genericity of the APs Framework by applying it to a different data source and domain of application.

Previously focused on X/Twitter and the tourism domain, we expanded our experimentation to include the new domain of local public policies. This involved using data from municipality review platforms throughout *France*, in collaboration with local public policy researchers. This shift allowed us to demonstrate how each step of the framework could be adapted to a new dataset and domain, thereby confirming its generic applicability. However, the transition to using data characterized by longer texts than typical social media posts brought to light new challenges, which could be explored in future extensions of our work.

Next, we will present a selection of perspectives aimed at extending and enhancing the proposals made in this thesis.

7.2 Perspectives

We will now explore future directions and perspectives that could enhance the proposals presented in this thesis. These perspectives not only aim to extend the current experiments but also to explore new avenues for applying the developed methodologies and tools in broader contexts.

[3] https://maxime-masson.github.io/TextBI

[1] https://www.geodatadays.fr/page/GeoDataDays-2023-Les-Challenges-Geodata/139

7.2.1 Perspective 1: Enhancing the APs Framework Experiments

The first perspective differs from the following ones, as it is strictly related to our experiments. We propose three major avenues to address their limitations.

Experimenting with the Framework on Massive Heterogeneous Datasets

A key aspect we have not yet explored is experimenting with the framework on massive datasets, on the order of hundreds of thousands of posts, potentially aggregated from various social media platforms. This is currently one of the main limitations of our experiments.

- For Contribution 1, this would involve assessing whether our collection methodology can produce accurate, larger-scale datasets. We anticipate new challenges, particularly with the iterative feedback loop mechanism, which might necessitate automation.
- For Contribution 2, it is essential to determine if our findings regarding the best NLP techniques and optimal number of annotated examples for the three knowledge extraction tasks (Sentiment Analysis, NER for Locations, and Fine-grained Thematic Concept Extraction) can be generalized to very large datasets (e.g., hundreds of thousands of posts) from various social media sources, with potential variations in content length and writing style.
- For Contribution 3, this task would involve assessing the computational costs of the formulas used in the proxemic similarity toolkit (*ProxMetrics*) to evaluate the performance required for rapid, real-time indicators on massive datasets.
- Lastly, for Contribution 4, it is crucial to observe how the *TextBI* platform performs at scale. Given it is a web application, some modifications may be necessary. Kelleher and Braswell (2021) offers advice on visualizing large-scale datasets that could be integrated into the *TextBI* platform, for example on the use of aggregation to emphasize distinctive patterns.

Comparing NLP Experiments Findings Across Various Domains

Currently, Contribution 2 focuses exclusively on the *tourism* domain, primarily due to time constraints. Extending these experiments to a broad range of domains could provide insightful comparisons. This involves performing similar experiments with unchanged parameters across different domains to identify strategies (e.g., rule-based, fine-tuning, few-shot prompting, etc.) that maintain consistent performance.

It is also crucial to assess if the number of annotated examples deemed adequate in the tourism domain leads to equally satisfactory outcomes in other domains, or if there are significant discrepancies. Identifying and understanding the causes of any variations are essential steps towards adapting NLP techniques to various application domains effectively.

Furthermore, deep learning and NLP are rapidly evolving fields. Since our experiments, new LLMs such as *Google's Gemini* (Team et al., 2023) and *Microsoft's Phi-2* (Mojan and Sébastien, 2023) have emerged and could be included in further experiments.

Experimenting with Automatic Semantic Resource Generation for Multi-Domain Analyses

As of now, each domain of application (such as *tourism*, *local public policies*, etc.) is analyzed separately in the APs Framework. As a reminder, these domains are modeled through semantic resources (e.g., dictionary, thesaurus, ontology). The semantic resource corresponding to the domain of interest is provided as input to the framework, and the quality of the analyses produced heavily depends on it.

However, creating a semantic resource is a time-consuming process, and some end users, such as social media researchers, may not be interested in a specific domain. Instead, they may seek to understand broader trends in social media without focusing on any particular domain. They might want to explore variations across seasons, geographical areas, and other dimensions. Our current approach, which relies on domain-specific semantic resources, does not support this type of analysis. This is a limitation.

We hypothesize that topic modeling techniques like BERTopic (Grootendorst, 2022) or Top2Vec (Angelov, 2020) could overcome these limitations. Designing a *sandbox* mode for our framework could be beneficial. In this mode, the only input would be a social media corpus, and the semantic resource would be automatically generated using topic modeling techniques, as described in recent studies by Oba et al. (2021) and de Boer et al. (2023). This approach would allow for more flexible and dynamic analyses, accommodating the requirements of users interested in general social media trends.

7.2.2 Perspective 2: Widening Analytical Dimensions of the APs Framework

The second perspective is linked to all phases of the framework. We noticed that informational entities of different types, such as *people*, *organizations*, *currencies*, *amounts*, *temperatures*, etc., are also present in social media posts, which we currently do not leverage for our analysis. Nadeau (2007) has identified over 100 informational entity types present in texts (including animals, characters, food, sports teams, etc.). These entities are useful in many application domains; for example, *currencies* are crucial to the financial domains (Fatum and Yamamoto, 2016), *amounts* are significant for the healthcare domain (e.g., related to drug dosage (Brown et al., 2005)), and *temperatures* influence tourism (e.g., visitor behaviors (Scott et al., 2008), hospitality electricity consumption (Pablo-Romero et al., 2019)). Currently, some of these entities are included in our thematic dimension, but, as reflected above, some application domains heavily rely on them, and our current framework therefore would be limited regarding these domains.

Addressing these dimensions would involve expanding the analytical capabilities of the APs Framework. We could hypothesize that by incorporating and extending the 5W1H model (Wang et al., 2010a) into our process, this model would provide a structured approach to extending existing dimensions and adding new ones, thereby allowing for more nuanced analyses of social media content. The 5W1H model is based on the six interrogative words: *Who*, *What*, *When*, *Where*, *Why*, and *How* (Wang et al., 2010a). It has been widely used in journalism (Keith et al., 2020; Wang, 2012), health diagnosis (Almeida et al., 2020; Little et al., 2023), and more generally in question answering systems (Wang and Chua, 2010). We already incorporate the *When* (temporal dimension), *Where* (spatial dimension), and *What* (thematic dimension), and to some extent the *Who* (personal dimension). However, we could further expand our analysis with additional dimensions such as the *How* (methodological dimension). The *Who* could also be used to deeply analyze the links between social media users through user mentions in their posts, a facet we have not explored in this thesis. Additionally, the *Why* could be used to model the reasoning, context, motivation, and implications behind social media posts. We could also extend the 5W1H structure to make it multilevel for each interrogative word. For example, the *What* could be subdivided into: *what amount*, *what currency*, *what weather*, the latter could then be further subdivided into *what temperature*, *what humidity*, *what cloud coverage*, etc. This structure could be described as an interrogative ontology, like in Kim et al. (2012). It would allow to have more granular analyses for domain stakeholders using simple interrogative words.

Incorporating the 5W1H model into our framework would imply focusing on designing algorithms and methodologies capable of automatically extracting and analyzing these dimensions from text. This adaptation would allow the framework to potentially address more complex requirements from domain stakeholders. Lastly, regarding the visualization aspect (*TextBI*), enhancing the modularity of the dashboard would be necessary so that it can handle more dimensions. We will address that in the next perspective.

7.2.3 Perspective 3: Industrializing the *TextBI* Dashboard

An important aspect related to Contribution 4 is the industrialization of the *TextBI* dashboard. Indeed, the dashboard was highly appreciated, especially at the *GeoDataDays 2023*, and it would be regrettable not to make it publicly available for reuse in other projects. However, further work is required. This industrialization is also related to Subsubsection 7.2.1 as it is essential to understand how the platform performs at scale with very large datasets.

Improvement to Graphical User Interfaces (GUI) and Colors

The first weakness of the *TextBI* dashboard is related to its graphical user interface (GUI) and the colors used in visuals. Currently, modifications in the granularity of visuals, such as adjustments in spatial, temporal, and thematic dimensions, require direct edits to a configuration file followed by a subsequent reload of the application. This is also required for loading different datasets. Thus, there is a requirement for dedicated and intuitive interfaces that allow users, especially those without extensive computer science backgrounds, to easily manage these adjustments.

Several studies (Bao et al., 2022; Li et al., 2018) have addressed the challenge of presenting data at varying levels of granularity to end-users, providing a potential foundation upon which our interfaces could be developed. This issue also affects the proxemic view (refer to Subsection 5.3.4). Currently, the adjustment of proxemic dimensions relies on sliders, a method that has been found challenging for novice users to interact with (e.g., understanding the specific effect of each slider). Enhancing this view by leveraging the eight golden rules of interface design proposed by Shneiderman (2004) could significantly improve the user experience, especially the principles of "*Cater to universal usability*" and "*Offer informative feedback*". However, proxemic dimensions encompass a range of complexities, and visually explicating their impacts without directly presenting the results poses a significant challenge. Existing research works have proposed approaches to display informative feedback to novice users, such as in the context of shopping websites (Chen and Zhai, 2023) or mobile applications (Punchoojit et al., 2017) that could be adapted for *TextBI*.

Ultimately, a thorough examination of the color schemes used in the dashboard is essential, particularly in elements such as map gradients, timeline visuals, and the coloration of proxemic entities. Feedback from end-users indicates that these aspects are not always intuitively comprehensible, highlighting the requirement for improvement. Existing works have examined the influence of colors on users' perception of data and explored the reasoning behind people's choices of color palettes (Ahmad et al., 2021; Szafir, 2017). The findings from these studies could serve as guidelines for integrating more intuitive colors into *TextBI*.

Enhancing Dashboard Customizability through End-User Modifications

Another limitation is the lack of customizability of the dashboard views. Currently, the layout of the screens is fixed (e.g., each visual will always be in the same position). This is a significant strain because different domains may not necessarily require all visuals to be used, and the hierarchy of visuals may not be the same for each domain. For example, in the domain of *urban planning*, the spatial dimension (mapping of resources or infrastructure) and temporal dimension (changes over time) are likely to be more important, as reflected by existing works in this domain (Isinkaralar, 2023). Additionally, if we extend our framework with additional analytical dimensions (refer to Subsection 7.2.2), this customizability will become mandatory to avoid cluttered views.

We therefore aim to make the platform modular and customizable by end-users. This could involve having a library of visualization modules (*e.g., thematic maps, timelines, association graphs, etc.*) and making them draggable onto the main screen to customize views. Various existing dashboards have implemented user customization, like Filonik et al. (2013) in the context of environmental performance visualizations leveraging the concept of *widgets* and Roberts et al. (2017) in the context of higher education. Vázquez-Ingelmo et al. (2019) provides a thorough review of the approaches used to build modular dashboards. This system would also allow for the creation of many more visualization modules. For example, we are considering alternative visuals for trajectories, such as the metro map-like visualizations proposed in Jacobsen et al. (2020). These metro map-like visualizations are specifically tailored for set systems, simplifying the depiction of social media users' thematic associations and sequences.

Integrating the APs Framework into a Unified Platform for Live Data Streams

Currently, *TextBI* is able to visualize any social media dataset which respects the APs Trajectory Model. The latter is instantiated using modules corresponding to the preceding phases of the APs Framework (e.g., *collection*, *transformation*). The data collection from social media and model instantiation have to be prepared beforehand, separately, and then given as input to *TextBI*. This architecture, therefore, does not directly support live data streams from social media.

Unifying all phases of the APs Framework in a single software platform would make it possible to support a real-time mode to track the evolution of themes, their geographical distribution, etc., during major events such as elections, concerts, or sports competitions (one of the most frequent requests we collected from end-users). It would also necessitate making all phases of the framework real-time, which introduces a lot of challenges and complexities (e.g., spatio-temporal computation delays, scalability, real-time implementation of deep learning models, etc.) (Mehmood and Anees, 2020). For instance, this unified platform could leverage live databases like *InfluxDB* (Ahmad and Ansari, 2017) to store and update the data model. Existing work have studied real-time post extraction and processing using X/Twitter data streams (Gupta and Hewett, 2020; Mazoyer et al., 2018) or distributed architectures (Efstathiades et al., 2016) and could be extended for use in our platform. This extension was not conducted as part of the thesis due to its time-consuming nature and the fact that it primarily involves engineering tasks.

7.2.4 Perspective 4: Extending Proxemic Applications

The final perspective proposes exploring new avenues in our use of *proxemics*. We start by addressing the main limitation identified: the difficulties encountered in processing data from platforms with longer texts, such as municipality review platforms. We propose ways to alleviate this issue so that our proxemic framework can better accommodate longer texts. This is related to all aspects of Contribution 3.

Next, we propose several new applications for the proxemic similarity toolkit (*ProxMetrics*, specifically Contribution 3.3). This toolkit, which interprets social media interactions through the lens of proxemics theory, could be extended and adapted for various other purposes beyond calculating domain-specific indicators (refer to Figure 1.5, *Valorize*).

Refining Analysis of Longer Texts Like Municipality Reviews

In Chapter 6, we applied the APs Trajectory Model and the *ProxMetrics* toolkit to analyze data sourced from municipality review platforms, which can, to an extent, be considered a kind of social media. Unlike traditional social media platforms such as X/Twitter, which typically host concise posts, municipality review platforms contain extensive, detailed reviews. These longer reviews often encapsulate a variety of themes, each expressing different sentiments, which introduces complexity into the sentiment analysis process.

We observed that in these reviews, citizens frequently express mixed feelings, praising certain aspects of local public policies while criticizing others within the same review. This multifaceted expression makes sentiment analysis challenging, as reviews cannot be simply classified into ternary sentiments like tweets often are. Due to these complexities, our current approach of attributing a uniform sentiment per post proves to be inadequate for platforms with longer texts. This method was promising on platforms like X/Twitter, where posts are brief and sentiments are generally more consistent. However, it significantly misrepresents the nuanced sentiments expressed in municipality reviews.

To more accurately capture the diverse sentiments expressed in reviews, we propose experimenting with segmenting posts into specific aspects using aspect-based sentiment analysis (ABSA) approaches (Brauwers and Frasincar, 2022; Nazir et al., 2020). ABSA is an advanced sentiment analysis technique that divides text into smaller segments, each corresponding to different features being discussed, and analyzes the sentiment for each aspect independently. This approach is particularly suited for addressing the complex expressions found in lengthy municipality reviews or blog posts. Various models referenced in Chapter 3, such as MLMs like BERT (Hoang et al., 2019) and LLMs like LLaMA 2 (Yang et al., 2024), have demonstrated promising results for ABSA in multi-domain contexts. Beyond the NLP aspect, we also need to refine the APs Trajectory Model to accommodate post segmentation. Multi-aspect semantic trajectory models, common in geomatics (Cayèré et al., 2021; Mello et al., 2019), could inspire extensions of our proxemic model to handle segmented posts. Lastly, we expect changes in how proxemic similarity is calculated to accommodate this new model; new formulas should be experimented with for each proxemic dimension.

Proxemic Recommender System

Another application of proxemic could be to leverage the proxemic similarity measurements produced by *ProxMetrics* as inputs for a domain-adaptable and modular recommender system. The hypothesis is that this would contribute to addressing the research challenge of proposing modular, domain-adaptive recommendations, customizable by non-computer scientist users. This is currently an active research topic, see Hao et al. (2024); Zang et al. (2022); Zhu et al. (2021).

The *ProxMetrics* toolkit is capable of computing similarities among various social media entities, including users, groups, themes, places, or periods, with support for heterogeneous combinations. Consequently, it would be natural to extend the work by proposing a generic recommender system capable of offering recommendations across multiple application domains dynamically. For instance, in the tourism domain, this could involve recommending touristic activities or Points of Interest (POIs) as well as creating a user recommender system to connect visitors with shared interests. Another major challenge would be to achieve *ubiquitous* proxemic similarity measurements, this means being able to recommend linked entities (e.g., a leisure activity associated with a particular place, weather, and period of the day).

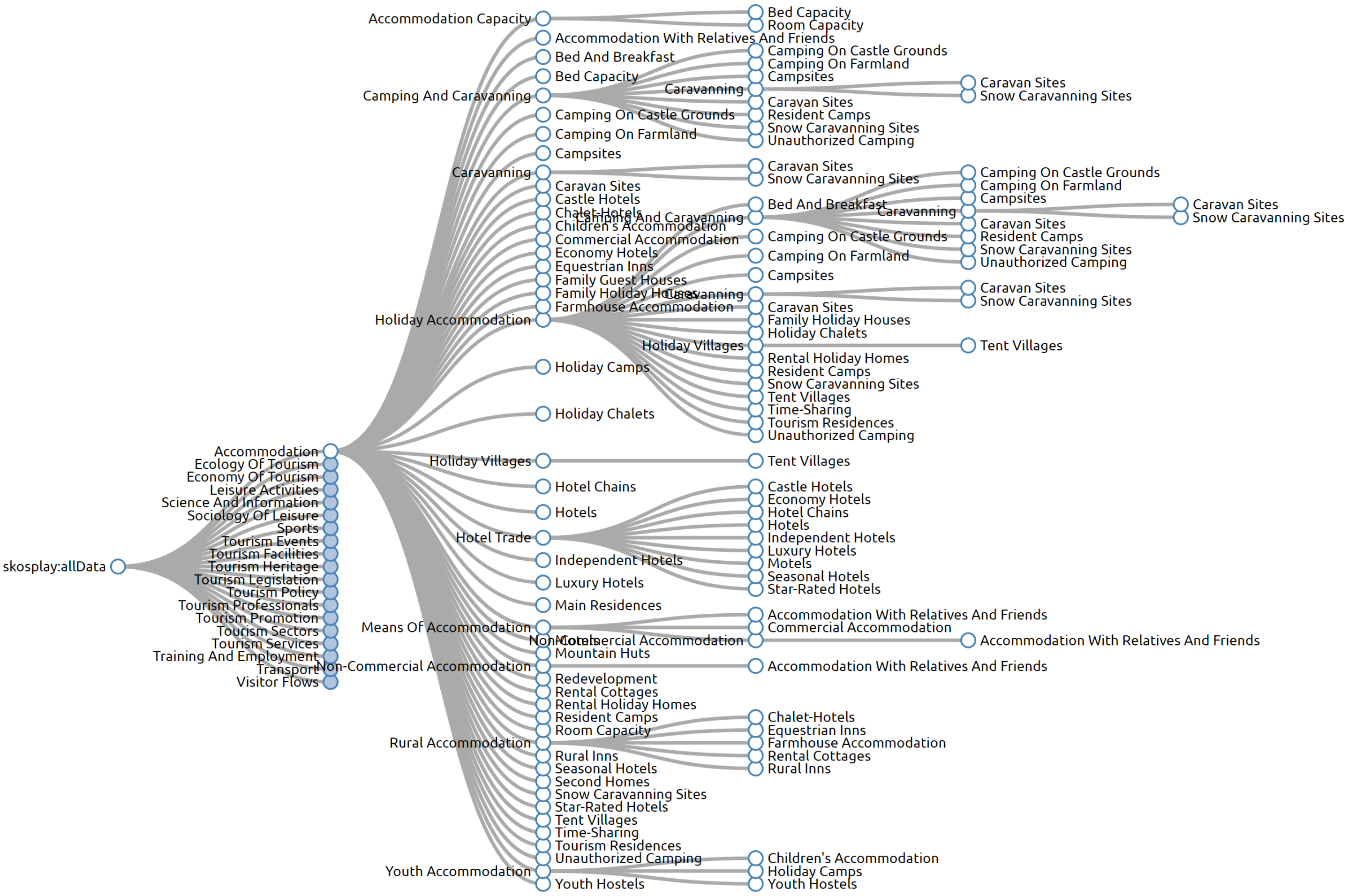
Integration of Proxemic Dimensions into a Domain-Specific Language for Social Media Querying

Lastly, we envision the integration of proxemic dimensions (Distance, Identity, Location, Movement, and Orientation) into a domain-specific language (DSL) tailored for querying social media platforms. This proxemic approach aims to extend the framework presented by Butakov et al. (2018), which provides a comprehensive, unified DSL for the collection and processing of social media data. The existing language, however, may pose usability challenges for individuals without a background in computer science. In contrast, a proxemic-enhanced DSL would enable users to perform queries on social media entities by leveraging proxemic dimensions as *metaphors*. Given that proxemic dimensions are inherently linked to the physical world, they represent a valuable asset for users lacking expertise in computer science, facilitating their ability to construct queries within social media datasets. This approach significantly diverges from conventional query languages (such as *SQL* (Date, 1984), *GraphQL* (Quiña-Mera et al., 2023), and *Cypher* (Francis et al., 2018)), which frequently present a steep learning curve to users unfamiliar with computer science terminology and concepts. The development of a proxemic query language introduces unique challenges, particularly in accurately mapping complex social interactions onto proxemic dimensions.

Appendices

Appendix A

Accommodation Branch of the Thesaurus on Tourism and Leisure Activities of the World Tourism Organization



Appendix B

Extracts of our Annotated Dataset

```
neutral    Journee a Saint-Jean-de-Luz. Petit cafe en terrasse pour commencer. #coffee
```

```
journee    B-Days
a    O
saint    O
-   O
jean    O
-    O
de    O
-    O
luz    O
.    O
petit    O
cafe    B-Cafes
en    O
terrasse    B-Terraces
pour    O
commencer    O
.    O
#coffee    B-Cafes
```

```
journee    O
a    O
saint    B-LOC
-   I-LOC
jean    I-LOC
-    I-LOC
de    I-LOC
-    I-LOC
luz    I-LOC
.    O
petit    O
cafe    O
en    O
```

```
terrasse    O
pour    O
commencer    O
.    O
#coffee O
```

Appendix C

Instantiation of the APs Trajectory Model in JSON

```json
{
    "source": "twitter",
    "groups": [
        {
            "id": "group_influencers",
            "groupName": "Influencers",
            "criteria": [
                {
                    "description": "followers > 5000"
                }
            ]
        }
    ],
    "identities": [
        {
            "type": "user",
            "id": "user_1",
            "name": "John",
            "features": [
                {
                    "key": "followers_count",
                    "value": 12581
                }
            ]
        },
        {
            "type": "user",
            "id": "user_2",
            "name": "Pierre",
            "features": [
                {
                    "key": "followers_count",
                    "value": 5484
                },
                {
                    "key": "country",
                    "value": "france"
                }
            ],
            "movement": [
                {
                    "post_id": "post_1",
                    "post_text": "Perfect weather, perfect water...
                       Wow :) at Camping Eskualduna",
                    "post_metadata": {
                        "timestamp": "2019-06-24T14:56:33.000Z"
                    },
                    "locations": [
                        {
                            "type": "time",
                            "fromMetadata": true,
                            "timestamp": "2019-06-24T14:56:33.000Z"
                        },
                        {
                            "type": "place",
                            "fromMetadata": false,
                            "startIndex": 44,
                            "length": 18,
                            "placeName": "Camping Eskualduna",
                            "uri": "https://www.osm.org/way/71565094"
                        },
                        {
                            "type": "theme",
                            "fromMetadata": false,
                            "startIndex": 25,
                            "length": 5,
                            "conceptName": "Weather ",
                            "uri": "Tourism://ClimaticFactor/Weather"
                        },
                        {
                            "type": "theme",
                            "fromMetadata": false,
                            "startIndex": 8,
                            "length": 7,
                            "conceptName": "Water",
                            "uri": "Tourism://NaturalResources/Water"
                        }
                    ],
                    "orientations": [
                        {
                            "type": "sentiment",
                            "polarity": "positive"
                        },
                        {
                            "type": "engagement",
                            "value": [
                                {
                                    "type": "likes",
                                    "value": 44
                                },
                                {
                                    "type": "reposts",
                                    "value": 2
                                },
                                {
                                    "type": "replies",
                                    "value": 5
                                },
                                {
                                    "type": "quotes",
                                    "value": 0
                                }
                            ]
                        }
                    ]
                },
                {
                    "post_id": "post_2",
                    "post_text": "what a pleasure to discover this beautiful Saint Vincent d’Hendaye Church while going to the market :)",
                    "post_metadata": {
                        "timestamp": "2019-06-29T18:12:15.000Z",
                        "geotag": {
                            "country": "France",
                            "name": "Biarritz",
                            "full_name": "Biarritz, France",
                            "country_code": "FR",
                            "place_type": "municipality"
                        }
                    },
                    "locations": [
                        {
                            "type": "time",
                            "fromMetadata": true,
                            "timestamp": "2019-06-29T18:12:15.000Z"
                        },
                        {
                            "type": "place",
                            "fromMetadata": true,
                            "startIndex": null,
                            "length": null,
                            "placeName": "Biarritz",
                            "uri": "https://www.osm.org/relation/4615703"
                        },
                        {
                            "type": "place",
                            "fromMetadata": false,
                            "startIndex": 52,
                            "length": 30,
                            "conceptName": "Saint Vincent d’Hendaye Church",
                            "uri": "https://osm.org/way/72070289"
                        },
                        {
                            "type": "theme",
                            "fromMetadata": false,
                            "startIndex": 102,
                            "length": 6,
                            "conceptName": "Market",
                            "uri": "Tourism://Commercial/Market"
                        }
                    ],
                    "orientations": [
                        {
                            "type": "sentiment",
                            "polarity": "positive"
                        },
                        {
                            "type": "engagement",
                            "value": [
                                {
                                    "type": "likes",
                                    "value": 15
                                },
                                {
                                    "type": "reposts",
                                    "value": 0
                                },
                                {
                                    "type": "replies",
                                    "value": 3
                                },
                                {
                                    "type": "quotes",
                                    "value": 0
                                }
                            ]
                        }
                    ]
                }
            ]
        }
    ],
    "distances": [
        {
            "type": "harversine",
            "from": "user_1",
            "to": "user_2",
            "value": 16,
            "unit": "kilometers"
        },
        {
            "type": "harversine",
            "from": "https://osm.org/way/72070289",
            "to": "https://osm.org/relation/4615703",
            "value": 23,
            "unit": "kilometers"
        },
        {
            "type": "semantic",
            "from": "Tourism://NaturalResources/Water",
            "to": "Tourism://ClimaticFactor/Weather",
            "value": 0.23,
            "unit": "ratio"
        },
        {
            "type": "temporal",
            "from": "2019-06-24 14:56:33",
            "to": " 2019-06-29 18:12:15",
            "value": 5,
            "unit": "days"
        }
    ]
}
```

Appendix D

Qualitative Evaluation Survey for *TextBI*

				Tourism Office	Public Policy Researcher
Target	**Criteria**	**Sub-Criteria**	**ID**	**Agreement**	**Agreement**
Frequency View	**Feel**	Intuitiveness	C1	●●●●○	●●●●○
		Feedback	C2	●●●●●	●●●●●
	Usability	Learnability	C3	●●●●●	●●●●○
		Memorability	C4	●●●●●	●●●●○
		Ease of use	C5	●●○○○	●●●●○
		Efficiency	C6	●●●●●	●●●●○
		Recoverability	C7	●●●●●	●●●●●
		Error Handling	C8	●●●○○	●●●○○
	Design	Global Aesthetics	C9	●●●●●	●●●●●
		Consistency	C10	●●●●●	●●●●○
		Colors	C11	●●●●○	●●○○○
		Clarity	C12	●●○○○	●●●●○
		Layout	C13	●●●●●	●●●●○
	Features	Completeness	C14	●○○○○	●●○○○
		Flexibility	C15	●●●●●	●●●○○
		Potential	C16	●●●●●	●●●●●
		Usefulness	C17	●●○○○	●●●●○
		Accuracy	C18	●●●●●	●○○○○
		Innovativeness	C19	●●●●●	●●●●●
		Compatibility	C20	●●●●●	●●●○○
	Performance	Reactiveness	C21	●●●●●	●●●●○
		Stability	C22	●●●●●	●●●●●
Association View	**Feel**	Intuitiveness	C1	●●●●○	●●●●○
		Feedback	C2	●●●●●	●●●●●
	Usability	Learnability	C3	●●○○○	●●●●○
		Memorability	C4	●●●●●	●●●●○
		Ease of use	C5	●●○○○	●●●●○
		Efficiency	C6	●●●●○	●●●●○
		Recoverability	C7	●●●●●	●●●●●
		Error Handling	C8	●●●○○	●●●○○
	Design	Global Aesthetics	C9	●●●●●	●●●●●
		Consistency	C10	●●●●●	●●●●○
		Colors	C11	●●●●●	●●●●○
		Clarity	C12	●●○○○	●●○○○
		Layout	C13	●●●●●	●●●●○

	Features	Completeness	C14	●○○○○	●●○○○
		Flexibility	C15	●●●●●	●●●○○
		Potential	C16	●●●●●	●●●●○
		Usefulness	C17	●●○○○	●●●●○
		Accuracy	C18	●●●●○	●○○○○
		Innovativeness	C19	●●●●●	●●●●○
		Compatibility	C20	●●●●●	●●●○○
	Performance	Reactiveness	C21	●●●●●	●●●●○
		Stability	C22	●●●●●	●●●●●
Movement View	**Feel**	Intuitiveness	C1	●●●●○	●●●●○
		Feedback	C2	●●●●●	●●●●●
	Usability	Learnability	C3	●●○○○	●●●●○
		Memorability	C4	●●●●●	●●●●○
		Ease of use	C5	●●○○○	●●○○○
		Efficiency	C6	●●○○○	●●●●○
		Recoverability	C7	●●●●●	●●●●○
		Error Handling	C8	●●●○○	●●●○○
	Design	Global Aesthetics	C9	●●●●●	●●●●○
		Consistency	C10	●●●●●	●●●●○
		Colors	C11	●●●●●	●●○○○
		Clarity	C12	●○○○○	●○○○○
		Layout	C13	●●○○○	●●●○○
	Features	Completeness	C14	●○○○○	●●○○○
		Flexibility	C15	●●●○○	●●●○○
		Potential	C16	●●●○○	●●●○○
		Usefulness	C17	●○○○○	●●●○○
		Accuracy	C18	●●●●○	●○○○○
		Innovativeness	C19	●●○○○	●●●○○
		Compatibility	C20	●●●●●	●●●○○
	Performance	Reactiveness	C21	●●●●●	●●○○○
		Stability	C22	●●●●●	●●●●●
***Proxemics* View**	**Feel**	Intuitiveness	C1	●●●●●	●●●●○
		Feedback	C2	●●●●●	●●●●●
	Usability	Learnability	C3	●●●●●	●●●●○
		Memorability	C4	●●●●●	●●●●○
		Ease of use	C5	●●○○○	●●○○○
		Efficiency	C6	●●●●●	●●●●○
		Recoverability	C7	●●●○○	●●●●●
		Error Handling	C8	●●●○○	●●●○○
	Design	Global Aesthetics	C9	●●○○○	●●●●●
		Consistency	C10	●●●●●	●●●●○
		Colors	C11	●●●●●	●●○○○
		Clarity	C12	●●●●●	●●●●○
		Layout	C13	●●●●●	●●●●○
	Features	Completeness	C14	●○○○○	●●○○○
		Flexibility	C15	●●●●●	●●●○○
		Potential	C16	●●●●●	●●●●●
		Usefulness	C17	●●○○○	●●●●○
		Accuracy	C18	●●●●●	●○○○○
		Innovativeness	C19	●●●●●	●●●●●
		Compatibility	C20	●●●●●	●●●●○
	Performance	Reactiveness	C21	●●●●●	●●●●○

		Stability	C22	●●●●●	●●●●●
Overlays	**Feel**	Intuitiveness	C1	●●●●●	●●●●○
		Feedback	C2	●●●●●	●●●●●
	Usability	Learnability	C3	●●●●●	●●●●○
		Memorability	C4	●●●●●	●●●●○
		Ease of use	C5	●●●●○	●●●●○
		Efficiency	C6	●●●●●	●●●●○
		Recoverability	C7	●●●●●	●●●●●
		Error Handling	C8	●●●○○	●●●○○
	Design	Global Aesthetics	C9	●●●●○	●●●●●
		Consistency	C10	●●●●●	●●●●○
		Colors	C11	●●●●○	●●●●●
		Clarity	C12	●●●●○	●●●●○
		Layout	C13	●●●○○	●●●●○
	Features	Completeness	C14	●○○○○	●●○○○
		Flexibility	C15	●●●●●	●●●○○
		Potential	C16	●●●●●	●●●●●
		Usefulness	C17	●●●●○	●●●●○
		Accuracy	C18	●●●●●	●○○○○
		Innovativeness	C19	●●●●●	●●●●●
		Compatibility	C20	●●●●●	●●●○○
	Performance	Reactiveness	C21	●●●●●	●●●●○
		Stability	C22	●●●●●	●●●●●
Global	**Feel**	Intuitiveness	C1	●●●●○	●●●●○
		Feedback	C2	●●●●●	●●●●●
	Usability	Learnability	C3	●●●●●	●●●●○
		Memorability	C4	●●●●●	●●●●○
		Ease of use	C5	●●○○○	●●●●○
		Efficiency	C6	●●●●●	●●●●○
		Recoverability	C7	●●●●●	●●●●●
		Error Handling	C8	●●●○○	●●●○○
	Design	Global Aesthetics	C9	●●●●●	●●●●●
		Consistency	C10	●●●●●	●●●●○
		Colors	C11	●●●●○	●●●●○
		Clarity	C12	●●●●○	●●●●○
		Layout	C13	●●●●○	●●●●○
	Features	Completeness	C14	●●○○○	●●○○○
		Flexibility	C15	●●●●●	●●●○○
		Potential	C16	●●●●●	●●●●●
		Usefulness	C17	●●●●○	●●●●○
		Accuracy	C18	●●●●●	●●○○○
		Innovativeness	C19	●●●●●	●●●●●
		Compatibility	C20	●●●●●	●●●○○
	Performance	Reactiveness	C21	●●●●●	●●●●○
		Stability	C22	●●●●●	●●●●●

Bibliography

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agbalagba, O., Avwiri, G., and Ononugbo, C. (2016). Gis mapping of impact of industrial activities on the terrestrial background ionizing radiation levels of ughelli metropolis and its environs, nigeria. *Environmental Earth Sciences*, 75:1–10.

Agerri, R., Centeno, R., Espinosa, M., Landa, J. F., and Rodrigo, A. (2021). Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194.

Agüero-Torales, M. M., Salas, J. I. A., and López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, 107:107373.

Ahmad, J., Huynh, E., and Chevalier, F. (2021). When red means good, bad, or canada: exploring people's reasoning for choosing color palettes. In *2021 IEEE Visualization Conference (VIS)*, pages 56–60. IEEE.

Ahmad, K. and Ansari, M. (2017). Hands-on influxdb. In *NoSQL: Database for Storage and Retrieval of Data in Cloud*, pages 341–354. Chapman and Hall/CRC.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., and Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):1–33.

Akram, W. and Kumar, R. (2017). A study on positive and negative effects of social media on society. *International journal of computer sciences and engineering*, 5(10):351–354.

Al-Harbi, O., Jusoh, S., and Norwawi, N. M. (2017). Lexical disambiguation in natural language questions (nlqs). *arXiv preprint arXiv:1709.09250*.

Alalwan, A. A., Rana, N. P., Dwivedi, Y. K., and Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and informatics*, 34(7):1177–1190.

Aldhaheri, A. and Lee, J. (2017). Event detection on large social media using temporal analysis. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1–6. IEEE.

Alfattni, G., Peek, N., and Nenadic, G. (2020). Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of biomedical informatics*, 108:103488.

Algan, Y., Malgouyres, C., and Senik, C. (2020). Territoires, bien-être et politiques publiques. *Les notes du conseil d'analyse économique*, (1):1–12.

Almeida, D., Machado, D., Andrade, J. C., Mendo, S., Gomes, A. M., and Freitas, A. C. (2020). Evolving trends in next-generation probiotics: a 5w1h perspective. *Critical reviews in food science and nutrition*, 60(11):1783–1796.

Alt, H. and Godau, M. (1995). Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02):75–91.

Amir, S., Wallace, B. C., Lyu, H., Carvalho, P., and Silva, M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 703–712.

Anderson, J., Casas Saez, G., Anderson, K., Palen, L., and Morss, R. (2019). Incorporating context and location into social media analysis: A scalable, cloud-based approach for more powerful data science.

Andrews, G., Issakidis, C., Sanderson, K., Corry, J., and Lapsley, H. (2004). Utilising survey data to inform public policy: comparison of the cost-effectiveness of treatment of ten mental disorders. *The British Journal of Psychiatry*, 184(6):526–533.

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anstead, N. and O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 uk general election. *Journal of computer-mediated communication*, 20(2):204–220.

Artetxe, M., Goswami, V., Bhosale, S., Fan, A., and Zettlemoyer, L. (2023). Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499.

Artetxe, M., Labaka, G., and Agirre, E. (2020). Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684.

Ashraf, M. U., Rehman, M., Zahid, Q., Naqvi, M. H., and Ilyas, I. (2021). A survey on emotion detection from text in social media platforms. *Lahore Garrison University Research Journal of Computer Science and Information Technology*, 5(2):48–61.

Ashrafi, N., Kelleher, L., and Kuilboer, J.-P. (2014). The impact of business intelligence on healthcare delivery in the usa. *Interdisciplinary Journal of Information, Knowledge, and Management*, 9:117.

Ata, A. (2022). *Analyse de la chaine de valeur des politiques publiques: le cas de la performance de la politique d'insertion et de maintien dans l'emploi des personnes en situation de handicap dans la fonction publique*. PhD thesis, Pau.

Ata, A. and Carassus, D. (2023). Pour en finir avec l'arlésienne de la gestion de la performance: la proposition d'une mesure globale et opérationnelle de la performance publique par sa chaîne de valeur. *Télescope: Revue d'analyse comparée en administration publique*.

Atout France (2023). Synthèse et sources de données. Accessed: 2023-11-20.

Azzone, G. (2018). Big data and public policies: Opportunities and challenges. *Statistics & Probability Letters*, 136:116–120.

Baars, H. and Kemper, H.-G. (2008). Management support with structured and unstructured data—an integrated business intelligence framework. *Information systems management*, 25(2):132–148.

Bajwa, I. S. and Choudhary, M. A. (2006). A rule based system for speech language context understanding. *Journal of Donghua University (English Edition) Vol*, 23(06).

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Banko, M. and Moore, R. C. (2004). Part-of-speech tagging in context. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 556–561.

Bao, H., Wang, G., Li, S., and Liu, Q. (2022). Multi-granularity visual explanations for cnn. *Knowledge-Based Systems*, 253:109474.

Barbieri, F., Camacho-Collados, J., Anke, L. E., and Neves, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Barger, V., Peltier, J. W., and Schultz, D. E. (2016). Social media and consumer engagement: a review and research agenda. *Journal of Research in Interactive Marketing*, 10(4):268–287.

Barklamb, A. M., Molenaar, A., Brennan, L., Evans, S., Choong, J., Herron, E., Reid, M., and McCaffrey, T. A. (2020). Learning the language of social media: a comparison of engagement metrics and social media strategies used by food and nutrition-related social media accounts. *Nutrients*, 12(9):2839.

Barnes, J., Oberlaender, L., Troiano, E., Kutuzov, A., Buchmann, J., Agerri, R., Øvrelid, L., and Velldal, E. (2022). Semeval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295.

Barriere, V. and Balahur, A. (2020). Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 266–271. International Committee on Computational Linguistics.

Basile, V., Lai, M., and Sanguinetti, M. (2018). Long-term social media data collection at the university of turin. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 1–6. CEUR-WS.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362.

Bastien, C. and Scapin, D. (1993). *Ergonomic criteria for the evaluation of human-computer interfaces*. PhD thesis, Inria.

Baucom, E., Sanjari, A., Liu, X., and Chen, M. (2013). Mirroring the real world in social media: Twitter, geolocation, and sentiment analysis. In *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing*, pages 61–68.

Becker, H., Naaman, M., and Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 291–300.

Bell, C., Fausset, C., Farmer, S., Nguyen, J., Harley, L., and Fain, W. B. (2013). Examining social media use among older adults. In *Proceedings of the 24th ACM conference on hypertext and social media*, pages 158–163.

Bellagente, M., Tow, J., Mahan, D., Phung, D., Zhuravinskyi, M., Adithyan, R., Baicoianu, J., Brooks, B., Cooper, N., Datta, A., Lee, M., Mostaque, E., Pieler, M., Pinnaparju, N., Rocha, P., Saini, H., Teufel, H., Zanichelli, N., and Riquelme, C. (2024). Stable lm 2 1.6b technical report.

Bello, B. S., Inuwa-Dutse, I., and Heckel, R. (2019). Social media campaign strategies: Analysis of the 2019 nigerian elections. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 142–149. IEEE.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pages 49–62.

Bergman, J. N., Buxton, R. T., Lin, H.-Y., Lenda, M., Attinello, K., Hajdasz, A. C., Rivest, S. A., Tran Nguyen, T., Cooke, S. J., and Bennett, J. R. (2022). Evaluating the benefits and risks of social media for wildlife conservation. *Facets*, 7(1):360–397.

Bergroth, L., Hakonen, H., and Raita, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48. IEEE.

Bhor, H. N., Koul, T., Malviya, R., and Mundra, K. (2018). Digital media marketing using trend analysis on social media. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 1398–1400. IEEE.

Biehl, J. T., Czerwinski, M., Smith, G., and Robertson, G. G. (2007). Fastdash: a visual dashboard for fostering awareness in software teams. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1313–1322.

Bijarchian, A. and Ali, R. (2014). Usability elements as benchmarking criteria for enterprise architecture methodologies. *J Teknol Sciences Eng*, 68(2):45–48.

Birhane, A., Kasirzadeh, A., Leslie, D., and Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bolton, C. et al. (2010). *Logistic regression and its application in credit scoring*. Phd thesis, University of Pretoria.

Bookstein, A., Kulyukin, V. A., and Raita, T. (2002). Generalized hamming distance. *Information Retrieval*, 5:353–375.

Booth, B., Mitchell, A., et al. (2001). Getting started with arcgis.

Bouabdallaoui, I., Guerouate, F., Bouhaddour, S., Saadi, C., and Sbihi, M. (2022). Named entity recognition applied on moroccan tourism corpus. *Procedia Computer Science*, 198:373–378.

Boudaa, B., Figuir, D., Hammoudi, S., and mohamed Benslimane, S. (2021). Datatourist: A constraint-based recommender system using datatourisme ontology. *International Journal of Decision Support System Technology (IJDSST)*, 13(2):62–84.

Bowman, S., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Brando, C., Domingués, C., and Capeyron, M. (2016). Evaluation of ner systems for the recognition of place mentions in french thematic corpora. In *Proceedings of the 10th workshop on geographic information retrieval*, pages 1–10.

Brauwers, G. and Frasincar, F. (2022). A survey on aspect-based sentiment classification. *ACM Computing Surveys*, 55(4):1–37.

Bredikhina, N., Gupta, K., and Kunkel, T. (2023). Superboosting the athlete social media brand: Events as an opportunity for follower growth. *European Sport Management Quarterly*, 23(6):1819–1842.

Brown, S. A., Blozis, S. A., Kouzekanani, K., Garcia, A. A., Winchell, M., and Hanis, C. L. (2005). Dosage effects of diabetes self-management education for mexican americans: the starr county border health initiative. *Diabetes care*, 28(3):527–532.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Brun, G., Dominguès, C., and Van Damme, M.-D. (2015). Textomap: determining geographical window for texts. In *Proceedings of the 9th workshop on geographic information retrieval*, pages 1–2.

Buccafurri, F., Fotia, L., and Lax, G. (2012). Allowing continuous evaluation of citizen opinions through social networks. In *Advancing Democracy, Government and Governance: Joint International Conference on Electronic Government and the Information Systems Perspective, and Electronic Democracy, EGOVIS/EDEM 2012, Vienna, Austria, September 3-6, 2012. Proceedings 1*, pages 242–253. Springer.

Bucher, B., Hein, C., Raines, D., and Gouet Brunet, V. (2021). Towards culture-aware smart and sustainable cities: Integrating historical sources in spatial information infrastructures. *ISPRS International Journal of Geo-Information*, 10(9):588.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Butakov, N., Petrov, M., Mukhina, K., Nasonov, D., and Kovalchuk, S. (2018). Unified domain-specific language for collecting and processing data of social media. *Journal of Intelligent Information Systems*, 51:389–414.

Caldwell, D. R. (2008). An analysis of toponymic homonyms in gazetteers: Country-level duplicate names in the national geospatial-intelligence agency's geographic names data base.

Campan, A., Atnafu, T., Truta, T. M., and Nolan, J. (2018). Is data collection through twitter streaming api useful for academic research? In *2018 IEEE international conference on big data (big data)*, pages 3638–3643. IEEE.

Cao, C. and Caverlee, J. (2015). Detecting spam urls in social media via behavioral analysis. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37*, pages 703–714. Springer.

Carassus, D. (2020). *Le pilotage des politiques publiques locales: de la planification à l'évaluation*.

Cardaioli, M., Conti, M., Di Sorbo, A., Fabrizio, E., Laudanna, S., and Visaggio, C. A. (2021). It'sa matter of style: Detecting social bots through writing style consistency. In *2021 International Conference on Computer Communications and Networks (ICCCN)*, pages 1–9. IEEE.

Castañer, M., Camerino, O., Anguera, M. T., and Jonsson, G. K. (2013). Kinesics and proxemics communication of expert and novice pe teachers. *Quality & Quantity*, 47(4):1813–1829.

Cayèré, C., Sallaberry, C., Faucher, C., Bessagnet, M.-N., Roose, P., Masson, M., and Richard, J. (2021). Multi-level and multiple aspect semantic trajectory model: application to the tourism domain. *ISPRS International Journal of Geo-Information*, 10(9):592.

Cesare, N., Grant, C., and Nsoesie, E. O. (2017). Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, pages 1–25.

Chang, K.-T. (2008). *Introduction to geographic information systems*, volume 4. Mcgraw-hill Boston.

Chang, K.-T. (2016). Geographic information system. *International encyclopedia of geography: people, the earth, environment and technology*, pages 1–10.

Chantrapornchai, C. and Tunsakul, A. (2021). Information extraction on tourism domain using spacy and bert. *ECTI Transactions on Computer and Information Technology*, 15(1):108–122.

Charalabidis, Y. and Loukis, E. (2012). Participative public policy making through multiple social media platforms utilization. *International Journal of Electronic Government Research (IJEGR)*, 8(3):78–97.

Chauhan, P., Sharma, N., and Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, 12:2601–2627.

Chavoshi, N., Hamooni, H., and Mueen, A. (2017). Temporal patterns in bot activities. In *Proceedings of the 26th international conference on world wide web companion*, pages 1601–1606.

Chen, B., Chen, B., Gao, D., Chen, Q., Huo, C., Meng, X., Ren, W., and Zhou, Y. (2021). Transformer-based language model fine-tuning methods for covid-19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 83–92. Springer.

Chen, C.-H. and Zhai, W. (2023). The effects of dynamic prompt and background transparency of hover feedback design on the user interface of shopping websites. *Asia Pacific Journal of Marketing and Logistics*, 35(4):809–827.

Chen, J., Liu, Y., and Zou, M. (2016). Home location profiling for users in social media. *Information & Management*, 53(1):135–143.

Cheng, X., Wang, W., Bao, F., and Gao, G. (2020). Mtner: A corpus for mongolian tourism named entity recognition. In *Machine Translation: 16th China Conference, CCMT 2020, Hohhot, China, October 10-12, 2020, Revised Selected Papers 16*, pages 11–23. Springer.

Chiruzzo, L., Castro, S., and Rosá, A. (2020). Haha 2019 dataset: A corpus for humor analysis in spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5106–5112.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Cignarella, A. T., Lai, M., Bosco, C., Patti, V., Paolo, R., et al. (2020). Overview of the task on stance detection in italian tweets. In *EVALITA 2020 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–10. Ceur.

Clark, E. and Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual english. *Procedia-Social and Behavioral Sciences*, 27:2–11.

Coghetto, R. (2016). Chebyshev distance. *Formalized Mathematics*, 24(2):121–141.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Communauté Pays Basque (2021). État des lieux tourisme pays basque synthèse. Technical report, Communauté Pays Basque, Pays Basque. Accessed: 2024-02-24.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Constantinides, E., Carmen Alarcón del Amo, M., and Romero, C. L. (2010). Profiles of social networking sites users in the netherlands.

Cossin, S., Jouhet, V., Mougin, F., Diallo, G., and Thiessard, F. (2018). Iam at clef ehealth 2018: Concept annotation and coding in french death certificates. *arXiv preprint arXiv:1807.03674*.

Crawford, M. and Khoshgoftaar, T. M. (2021). Using inductive transfer learning to improve hotel review spam detection. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 248–254. IEEE.

Crickard III, P. (2014). *Leaflet. js essentials*. Packt Publishing Ltd.

Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., and Murino, V. (2011). Towards computational proxemics: Inferring social relations from interpersonal distances. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 290–297. IEEE.

Cunha, J., Duarte, R., Guimarães, T., and Santos, M. F. (2023). Openehr and business intelligence in healthcare: an overview. *Procedia Computer Science*, 220:874–879.

Cuzzocrea, A., De Maio, C., Fenza, G., Loia, V., and Parente, M. (2016). Olap analysis of multidimensional tweet streams for supporting advanced analytics. In *Proceedings of the 31st annual ACM symposium on applied computing*, pages 992–999.

Da Rocha, H. (2019). *Learn Chart. js: Create interactive visualizations for the web with chart. js 2*. Packt Publishing Ltd.

Date, C. (1984). A critique of the sql database language. *ACM Sigmod Record*, 14(3):8–54.

Datig, I. and Whiting, P. (2018). Telling your library story: tableau public for data visualization. *Library Hi Tech News*, 35(4):6–8.

de Boer, M. H., Bakker, R. M., and Burghoorn, M. (2023). Creating dynamically evolving ontologies: A use case from the labour market domain. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*.

Derczynski, L., Bontcheva, K., and Roberts, I. (2016). Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan.

Desai, Z., Anklesaria, K., and Balasubramaniam, H. (2021). Business intelligence visualization using deep learning based sentiment analysis on amazon review data. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dey, L., Haque, S. M., Khurdiya, A., and Shroff, G. (2011). Acquiring competitive intelligence from social media. In *Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data*, pages 1–9.

Di Minin, E., Tenkanen, H., and Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3:63.

Diamantini, C., Mircoli, A., and Potena, D. (2016). A negation handling technique for sentiment analysis. In *2016 international conference on collaboration technologies and systems (cts)*, pages 188–195. IEEE.

Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., Zheng, H., and Liu, Z. (2021). Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213.

Domingués, C. and Eshkol-Taravella, I. (2015). Toponym recognition in custom-made map titles. *International Journal of Cartography*, 1(1):109–120.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Duarte, J. M., Santos, J. B. d., and Melo, L. C. (1999). Comparison of similarity coefficients based on rapd markers in the common bean. *Genetics and Molecular Biology*, 22:427–432.

Efstathiades, H., Antoniades, D., Pallis, G., and Dikaiakos, M. D. (2016). Distributed large-scale data collection in online social networks. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pages 373–380. IEEE.

El Abaddi, A., Backstrom, L., Chakrabarti, S., Jaimes, A., Leskovec, J., and Tomkins, A. (2011). Social media: source of information or bunch of noise. In *Proceedings of the 20th International conference companion on World Wide Web*, pages 327–328.

Elias, C. M. A. L. (2022). Twitter observatory: developing tools to recover and classify information for the social network twitter. Master's thesis, Universidade do Minho (Portugal).

Enríquez, M. P., Mencía, J. A., and Segura-Bedmar, I. (2022). Transformers approach for sentiment analysis: Classification of mexican tourists reviews from tripadvisor. *IberLEF 2022*.

Esmaeili, L., Nasiri, M., and Minaei-Bidgoli, B. (2011). Personalizing group recommendation to social network users. In *Web Information Systems and Mining: International Conference, WISM 2011, Taiyuan, China, September 24-25, 2011, Proceedings, Part I*, pages 124–133. Springer.

Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M., and Soroa, A. (2024). Latxa: An open language model and evaluation suite for basque. *arXiv preprint arXiv:2403.20266*.

Evans, D., Bratton, S., and McKee, J. (2021). *Social media marketing*. AG Printing & Publishing.

Fanzo, J., Haddad, L., McLaren, R., Marshall, Q., Davis, C., Herforth, A., Jones, A., Beal, T., Tschirley, D., Bellows, A., et al. (2020). The food systems dashboard is a new tool to inform better food policy. *Nature Food*, 1(5):243–246.

Fatum, R. and Yamamoto, Y. (2016). Intra-safe haven currency behavior during the global financial crisis. *Journal of International Money and Finance*, 66:49–64.

Ferrara, E. (2023). Social bot detection in the age of chatgpt: Challenges and opportunities. *First Monday*.

Ferrari, A. and Russo, M. (2016). *Introducing Microsoft Power BI*. Microsoft Press.

Filonik, D., Medland, R., Foth, M., and Rittenbruch, M. (2013). A customisable dashboard display for environmental performance visualisations. In *Persuasive Technology: 8th International Conference, PERSUASIVE 2013, Sydney, NSW, Australia, April 3-5, 2013. Proceedings 8*, pages 51–62. Springer.

Fishkin, J. S. (2003). Consulting the public through deliberative polling. *Journal of policy analysis and management*, 22(1):128–133.

Flood, M. D., Lemieux, V. L., Varga, M., and Wong, B. W. (2016). The application of visual analytics to financial stability monitoring. *Journal of financial stability*, 27:180–197.

Floris, R., Campagna, M., et al. (2014). Social media data in tourism planning: analysing tourists' satisfaction in space and time. In *REAL CORP 2014. Plan it Smart. Clever Solutions for Smart CitiesProceedings of19th International Conference on Urban Planning, Regional Development and Information Society*, pages 997–1003. Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE,.

Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., and Taylor, A. (2018). Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 international conference on management of data*, pages 1433–1445.

Friburger, N. and Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93–104.

Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., Wen, A., et al. (2020). Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, 109:103526.

Gainous, J. and Wagner, K. M. (2014). *Tweeting to power: The social media revolution in American politics*. Oxford University Press.

Gaizauskas, R. and Humphreys, K. (1997). Using a semantic network for information extraction. *Natural Language Engineering*, 3(2):147–169.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., and Choukri, K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, pages 139–142. Citeseer.

García-Ferrero, I., Agerri, R., and Rigau, G. (2022). Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416. Association for Computational Linguistics.

García-Ferrero, I., Agerri, R., and Rigau, G. (2023). T-projection: High quality annotation projection for sequence labeling tasks. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

García-Ferrero, I., Agerri, R., Salazar, A. A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., Molinet, B., Ramirez-Romero, J., Rigau, G., et al. (2024). Medical mt5: An open-source multilingual text-to-text llm for the medical domain. *arXiv preprint arXiv:2404.07613*.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922.

Gasparetto, A., Marcuzzo, M., Zangari, A., and Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83.

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., Foote Jr, J., Moseley, E. T., Grant, D. W., Tyler, P. D., et al. (2017). Comparing rule-based and deep learning models for patient phenotyping. *arXiv preprint arXiv:1703.08705*.

Gerosa, A. and Ceinar, I. M. (2022). New working spaces and covid-19: Analyzing the debate through twitter. In *The COVID-19 Pandemic and the Future of Working Spaces*, pages 11–24. Routledge.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12):4492–4511.

Ghag, K. V. and Shah, K. (2016). Negation handling for sentiment classification. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–6. IEEE.

Gkikas, D. C., Tzafilkou, K., Theodoridis, P. K., Garmpis, A., and Gkikas, M. C. (2022). How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in facebook. *International Journal of Information Management Data Insights*, 2(1):100067.

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., and Dutta, R. (2016). The language of mental health problems in social media. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 63–73.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Goel, A. and Gupta, L. (2020). Social media in the times of covid-19. *Journal of clinical rheumatology*.

Goyal, A., Gupta, V., and Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

Graham, F. (2023). Daily briefing: What the end of twitter's free api means for research. *Nature*.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., and Shoor, I. (2015). Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems*, 9(4):407–417.

Greenberg, S., Marquardt, N., Ballendat, T., Diaz-Marino, R., and Wang, M. (2011). Proxemic interactions: the new ubicomp? *interactions*, 18(1):42–50.

Groenen, P. J., Mathar, R., and Heiser, W. J. (1995). The majorization approach to multidimensional scaling for minkowski distances. *Journal of Classification*, 12:3–19.

Grønli, T.-M., Ghinea, G., and Younas, M. (2014). Context-aware and automatic configuration of mobile devices in cloud-enabled ubiquitous computing. *Personal and ubiquitous computing*, 18:883–894.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? *Handbook on ontologies*, pages 1–17.

Gunawan, A. B., Pratama, B., and Sarwono, R. (2021). Digital proxemics approach in cyber space analysis—a systematic literature review. *ICIC Express Letters*, 15(2):201–208.

Gundecha, P. and Liu, H. (2012). Mining social media: a brief introduction. *New directions in informatics, optimization, logistics, and production*, pages 1–17.

Gupta, P. and Bagchi, A. (2024). Data visualization with python. In *Essentials of Python for Artificial Intelligence and Machine Learning*, pages 237–282. Springer.

Gupta, V. and Hewett, R. (2020). Real-time tweet analytics using hybrid hashtags on twitter big data streams. *Information*, 11(7):341.

Gutiérrez, B. J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., and Su, Y. (2022). Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Hall, E. T. (1966). *The hidden dimension*, volume 609. Anchor.

Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold Jr, A. R., Durbin, M., Edmonson, M. S., Fischer, J., Hymes, D., Kimball, S. T., et al. (1968). Proxemics [and comments and replies]. *Current anthropology*, 9(2/3):83–108.

Hamstead, Z. A., Fisher, D., Ilieva, R. T., Wood, S. A., McPhearson, T., and Kremer, P. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*, 72:38–50.

Han, Q., Nesi, P., Pantaleo, G., and Paoli, I. (2020). Smart city dashboards: design, development, and evaluation. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–4. IEEE.

Han, W., McCabe, S., Wang, Y., and Chong, A. Y. L. (2018). Evaluating user-generated content in social media: an effective approach to encourage greater pro-environmental behavior in tourism? *Journal of Sustainable Tourism*, 26(4):600–614.

Hans, A. and Hans, E. (2015). Kinesics, haptics and proxemics: Aspects of non-verbal communication. *IOSR Journal of Humanities and Social Science (IOSR-JHSS)*, 20(2):47–52.

Hansoti, B. (2010). Business intelligence dashboard in decision making. *Purdue University, College of Technology Directed Projects*.

Hao, B., Yang, C., Guo, L., Yu, J., and Yin, H. (2024). Motif-based prompt learning for universal cross-domain recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 257–265.

Harpe, S. E. (2015). How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850.

Hartmann, J., Heitmann, M., Siebert, C., and Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Hasyim, M. (2019). Linguistic functions of emoji in social media communication. *Opcion*, 35.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Healey and Ramaswamy (2022). Twitter sentiment visualization. Accessed: 2023-06-30.

Hoang, M., Bihorac, O. A., and Rouces, J. (2019). Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196.

Höchtl, J., Parycek, P., and Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2):147–169.

Holt, K., Shehata, A., Strömbäck, J., and Ljungberg, E. (2013). Age and the effects of news media attention and social media use on political interest and participation: Do social media function as leveller? *European journal of communication*, 28(1):19–34.

Höpken, W. and Fuchs, M. (2022). Business intelligence in tourism. In *Handbook of e-Tourism*, pages 497–527. Springer.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Howson, C., Newbould, E., Duey, C., and Stockwell, M. T. (2012). *SAP BusinessObjects BI 4.0: The Complete Reference*. McGraw-Hill/Osborne Media.

Huang, C. (2017). Time spent on social network sites and psychological well-being: A meta-analysis. *Cyberpsychology, Behavior, and Social Networking*, 20(6):346–354.

Huang, L., Liu, G., Chen, T., Yuan, H., Shi, P., and Miao, Y. (2021). Similarity-based emergency event detection in social media. *Journal of safety science and resilience*, 2(1):11–19.

Hub, M. and Zatloukal, M. (2008). Methodology of fuzzy usability evaluation of information systems in public administration. *WSEAS Transactions on Information Science & Applications*, 5(11):1573–1583.

Hudders, L. and De Jans, S. (2022). Gender effects in influencer marketing: an experimental study on the efficacy of endorsements by same-vs. other-gender social media influencers on instagram. *International Journal of Advertising*, 41(1):128–149.

Humphreys, L. (2013). Mobile social media: Future challenges and opportunities. *Mobile Media & Communication*, 1(1):20–25.

Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S., and Fernández, S. (2016). A twitter sentiment gold standard for the brexit referendum. In *Proceedings of the 12th international conference on semantic systems*, pages 193–196.

Hussain, M. N., Obadimu, A., Bandeli, K. K., Nooman, M., Al-khateeb, S., and Agarwal, N. (2017). A framework for blog data collection: challenges and opportunities. In *The IARIA international symposium on designing, validating, and using datasets (DATASETS 2017)*.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Hvass, K. A. and Munar, A. M. (2012). The takeoff of social media in tourism. *Journal of vacation marketing*, 18(2):93–103.

Iman, M., Arabnia, H. R., and Rasheed, K. (2023). A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40.

INSEE (2023a). Insee - statistiques locales. Accessed: 2023-11-20.

INSEE (2023b). Insee - tableau de bord de l'économie française. Accessed: 2023-11-20.

Intelligence, M. (2009). Business intelligence in manufacturing. *MAIA*.

Isinkaralar, O. (2023). Spatio-temporal patterns of climate parameter changes in western mediterranean basin of türkiye and implications for urban planning. *Air Quality, Atmosphere & Health*, 16(11):2351–2363.

Isère Attractivité (2023). Carnet observatoires. Accessed: 2023-11-20.

Izhar, T. A. T., Baharuddin, M. F., Mohamad, A. N., Ramli, A. A. M., Shoid, M., and Hasnol, W. M. H. W. (2016). Using ontology for goal-based query to evaluate social media data. *Journal of Advances in Humanities and Social Sciences*, 2(2):108–118.

Jacobsen, B., Wallinger, M., Kobourov, S., and Nöllenburg, M. (2020). Metrosets: Visualizing sets as metro maps. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1257–1267.

Jander, H., Borgvall, J., and Castor, M. (2011). Brain budget–evaluation of human machine interaction in system development for high risk and task critical environments. Technical report, FOI-R–3272–SE.

Jankowski, P. (2009). Towards participatory geographic information systems for community-based environmental decision making. *Journal of environmental management*, 90(6):1966–1971.

Jayawardhana, U. K. and Gorsevski, P. V. (2019). An ontology-based framework for extracting spatio-temporal influenza data using twitter. *International journal of digital earth*, 12(1):2–24.

Jebb, A. T., Ng, V., and Tay, L. (2021). A review of key likert scale development advances: 1995–2019. *Frontiers in psychology*, 12:637547.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiang, H., Hua, Y., Beeferman, D., and Roy, D. (2022). Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208.

Jiang, L. and Yang, C. C. (2017). User recommendation in healthcare social media by assessing user similarity in heterogeneous network. *Artificial intelligence in medicine*, 81:63–77.

Jiashun, C. (2012). A new trajectory clustering algorithm based on traclus. In *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*, pages 783–787. IEEE.

Johansson, F., Kaati, L., and Shrestha, A. (2013). Detecting multiple aliases in social media. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 1004–1011.

Jooste, C., Van Biljon, J., and Mentz, J. (2014). Usability evaluation for business intelligence applications: A user support perspective. *South African Computer Journal*, 53(si-1):32–44.

Joshi, A., Kale, S., Chandel, S., and Pal, D. K. (2015). Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Kadam, S. and Vaidya, V. (2020). Review and analysis of zero, one and few shot learning approaches. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 1*, pages 100–112. Springer.

Kamath, U., Liu, J., and Whitaker, J. (2019). *Deep learning for NLP and speech recognition*, volume 84. Springer.

Kang, J. and Lee, H. (2017). Modeling user interest in social media using news media and wikipedia. *Information Systems*, 65:52–64.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Keith, B., Horning, M., and Mitra, T. (2020). Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. *Computational Journalism C+ J*.

Kelleher, C. and Braswell, A. (2021). Introductory overview: Recommendations for approaching scientific visualization with large environmental datasets. *Environmental Modelling & Software*, 143:105113.

Kenwright, B. (2019). Visualization with three. js. In *12th ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia 2019*.

Khalifa, M. b., Diaz Redondo, R. P., Vilas, A. F., and Rodríguez, S. S. (2017). Identifying urban crowds using geo-located social media data: a twitter experiment in new york city. *Journal of Intelligent Information Systems*, 48:287–308.

Khan, M. U., Choi, J. P., Shin, H., and Kim, M. (2008). Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *2008 30th annual international conference of the IEEE engineering in medicine and biology society*, pages 5148–5151. IEEE.

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Kim, D.-J., Lee, H.-W., Jung, J.-H., and Yong, H.-y. (2015). Usability evaluation criteria of software gui on weapon system. *Journal of the Korea Academia-Industrial cooperation Society*, 16(12):8691–8699.

Kim, J. (2021). The meaning of numbers: Effect of social media engagement metrics in risk communication. *Communication Studies*, 72(2):195–213.

Kim, J.-D., Son, J., and Baik, D.-K. (2012). Ca 5w1h onto: ontological context-aware model based on 5w1h. *International Journal of Distributed Sensor Networks*, 8(3):247346.

Kinra, A., Beheshti-Kashi, S., Buch, R., Nielsen, T. A. S., and Pereira, F. (2020). Examining the potential of textual big data analytics for public policy decision-making: A case study with driverless cars in denmark. *Transport Policy*, 98:68–78.

Kitani, T., Eriguchi, Y., and Hara, M. (1994). Pattern matching and discourse processing in information extraction from japanese text. *Journal of Artificial Intelligence Research*, 2:89–110.

Knoll, J. (2016). Advertising in social media: a review of empirical evidence. *International journal of Advertising*, 35(2):266–300.

Kok, S. and Domingos, P. (2008). Extracting semantic networks from text via relational clustering. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I 19*, pages 624–639. Springer.

Kosmajac, D. and Keselj, V. (2019). Twitter bot detection using diversity measures. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 1–8.

Krumm, J., Davies, N., and Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4):10–11.

Kumar, A., Trueman, T. E., and Cambria, E. (2021). Fake news detection using xlnet fine-tuning model. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, pages 1–4. IEEE.

Kumar, K. N. and Vineela, K. (2020). Friend recommendation using graph mining on social media. *International Journal of Engineering Technology and Management sciences (IJETMS) ijetms. in*, 4(5).

Kurt Menke, G., Smith Jr, R., Pirelli, L., John Van Hoesen, G., et al. (2016). *Mastering QGIS*. Packt Publishing Ltd.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Lahitani, A. R., Permanasari, A. E., and Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6. IEEE.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Landa, J. F. and Agerri, R. (2021). Social analysis of young basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0:1–15.

Lawson, N., Eustice, K., Perkowitz, M., and Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 71–79.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490.

Lee, D., Hosanagar, K., and Nair, H. S. (2018). Advertising content and consumer engagement on social media: Evidence from facebook. *Management Science*, 64(11):5105–5131.

Leung, D., Law, R., Van Hoof, H., and Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of travel & tourism marketing*, 30(1-2):3–22.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks*, 30(4):330–342.

Leys, C., Klein, O., Dominicy, Y., and Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of experimental social psychology*, 74:150–156.

Li, M., Bao, Z., Choudhury, F., and Sellis, T. (2018). Supporting large-scale geographical visualization in a multi-granularity way. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 767–770.

Lieber, O., Sharir, O., Lenz, B., and Shoham, Y. (2021). Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 1:9.

Lim, E.-P., Chen, H., and Chen, G. (2013). Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems (TMIS)*, 3(4):1–10.

Little, D., Amadio, P. C., Awad, H. A., Cone, S. G., Dyment, N. A., Fisher, M. B., Huang, A. H., Koch, D. W., Kuntz, A. F., Madi, R., et al. (2023). Preclinical tendon and ligament models: Beyond the 3rs (replacement, reduction, and refinement) to 5w1h (why, who, what, where, when, how). *Journal of Orthopaedic Research®*, 41(10):2133–2162.

Liu, H., Chatterjee, I., Zhou, M., Lu, X. S., and Abusorrah, A. (2020). Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375.

Liu, H., Hu, Z., Mian, A., Tian, H., and Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-based systems*, 56:156–166.

Liu, H., Morstatter, F., Tang, J., and Zafarani, R. (2016). The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, 1:137–143.

Liu, H., Wu, S. T., Li, D., Jonnalagadda, S., Sohn, S., Wagholikar, K., Haug, P. J., Huff, S. M., and Chute, C. G. (2012). Towards a semantic lexicon for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2012, page 568. American Medical Informatics Association.

Liu, X., Chen, H., and Xia, W. (2022). Overview of named entity recognition. *Journal of Contemporary Educational Research*, 6(5):65–68.

Liu, Y., Liu, Z., Chua, T.-S., and Sun, M. (2015). Topical word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Llobera, J., Spanlang, B., Ruffini, G., and Slater, M. (2010). Proxemics with multiple dynamic characters in an immersive virtual environment. *ACM Trans. Appl. Percept.*, 8(1).

Lo, S. L., Cambria, E., Chiong, R., and Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48:499–527.

Loubère, L. and Ratinaud, P. (2014). Documentation iramuteq 0.6 alpha 3 version 0.1. *http://www.iramuteq.org*.

Lu, Y., Wang, R., Zhang, Y., Su, H., Wang, P., Jenkins, A., Ferrier, R. C., Bailey, M., and Squire, G. (2015). Ecosystem health towards sustainability. *Ecosystem Health and Sustainability*, 1(1):1–15.

Luxey, A. (2019). *E-squads: a novel paradigm to build privacy-preserving ubiquitous applications*. Phd thesis, Université Rennes 1.

Ma, R., Zhou, X., Gui, T., Tan, Y., Li, L., Zhang, Q., and Huang, X. (2022). Template-free prompt tuning for few-shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5721–5732. Association for Computational Linguistics.

Ma, Y., Yang, X., Liao, L., Cao, Y., and Chua, T.-S. (2019). Who, where, and what to wear? extracting fashion knowledge from social media. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 257–265.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Machálek, T. (2020). Kontext: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008.

Majid, A., Chen, L., Chen, G., Mirza, H. T., Hussain, I., and Woodward, J. (2013). A context-aware personalized travel recommendation system based on geotagged social media data mining. *International Journal of Geographical Information Science*, 27(4):662–684.

Malmasi, S., Fang, A., Fetahu, B., Kar, S., and Rokhlenko, O. (2022). Multiconer: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117:30046–30054.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Mariani, M., Baggio, R., Fuchs, M., and Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, 30(12):3514–3554.

Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue.

Martin, L., Muller, B., Suarez, P. O., Dupont, Y., Romary, L., De La Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Massaron, L. (2024a). Fine-tune Llama 2 for sentiment analysis — kaggle.com. https://www.kaggle.com/code/lucamassaron/fine-tune-llama-2-for-sentiment-analysis. [Accessed 20-04-2024].

Massaron, L. (2024b). Fine-tune Mistral v0.2 for sentiment analysis — kaggle.com. https://www.kaggle.com/code/lucamassaron/fine-tune-mistral-v0-2-for-sentiment-analysis. [Accessed 20-04-2024].

Masson, M. (2022). Services augmentés pour le tourisme intelligent et l'analyse des pratiques. In *Forum Jeunes Chercheuses Jeunes Chercheurs (JCJC)-INFORSID 2022*.

Masson, M., Abdelhedi, S., Sallaberry, C., Agerri, R., Bessagnet, M.-N., Le Parc-Lacayrelle, A., and Roose, P. (2023a). Visualisation interactive de trajectoires d'activités touristiques application à des données extraites de twitter. In *Atelier Exploration des traces dans un monde du tout numérique: enjeux et perspectives-INFORSID 2023*.

Masson, M., Agerri, R., Sallaberry, C., Bessagnet, M.-N., Lacayrelle, A. L. P., and Roose, P. (2024a). Stratégies optimales pour l'analyse multidimensionnelle de contenus multilingues issus des réseaux sociaux. In *INFORSID 2024*.

Masson, M., Roose, P., Sallaberry, C., Agerri, R., Bessagnet, M.-N., and Lacayrelle, A. L. P. (2023b). Aps: A proxemic framework for social media interactions modeling and analysis. In *Advances in Intelligent Data Analysis XXI: 21st International Symposium on Intelligent Data Analysis, IDA 2023, Louvain-la-Neuve, Belgium, April 12–14, 2023, Proceedings*, pages 287–299. Springer.

Masson, M., Roose, P., Sallaberry, C., Bessagnet, M.-N., Le Parc Lacayrelle, A., and Agerri, R. (2024b). Proxmetrics: modular proxemic similarity toolkit to generate domain-adaptable indicators from social media. *Social Network Analysis and Mining*, 14(1):1–23.

Masson, M., Sallaberry, C., Agerri, R., Bessagnet, M.-N., Roose, P., and Le Parc Lacayrelle, A. (2022). A domain-independent method for thematic dataset building from social media: The case of tourism on twitter. In *Web Information Systems Engineering–WISE 2022: 23rd International Conference, Biarritz, France, November 1–3, 2022, Proceedings*, pages 11–20. Springer.

Masson, M., Sallaberry, C., Bessagnet, M.-N., Lacayrelle, A. L. P., Roose, P., and Agerri, R. (2024c). Textbi: An interactive dashboard for visualizing multidimensional nlp annotations in social media data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 1–9.

Maylawati, D. S., Zulfikar, W. B., Slamet, C., Ramdhani, M. A., and Gerhana, Y. A. (2018). An improved of stemming algorithm for mining indonesian text with slang on social media. In *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–6. IEEE.

Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. In *Workshop Programme*, page 15.

Maynard, D., Cunningham, H., Bontcheva, K., Catizone, R., Demetriou, G., Gaizauskas, R., Hamza, O., Hepple, M., Herring, P., Mitchell, B., et al. (2000). A survey of uses of gate. Technical report, Technical Report CS–00–06, Department of Computer Science, University of Sheffield.

Mazhari, S., Fakhrahmad, S. M., and Sadeghbeygi, H. (2015). A user-profile-based friendship recommendation solution in social networks. *Journal of Information Science*, 41(3):284–295.

Mazoyer, B., Cagé, J., Hudelot, C., and Viaud, M.-L. (2018). Real-time collection of reliable and representative tweets datasets related to news events. In *First International Workshop on Analysis of Broad Dynamic Topics over Social Media (BroDyn 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018)*.

McArdle, G. and Kitchin, R. (2016). The dublin dashboard: Design and development of a real-time analytical urban dashboard. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:19–25.

McCall, C. (2015). Mapping social interactions: the science of proxemics. *Social behavior from rodents to humans*, pages 295–308.

McGregor, S. C. (2019). Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, 20(8):1070–1086.

McKenna, S., Staheli, D., Fulcher, C., and Meyer, M. (2016). Bubblenet: A cyber security dashboard for visualizing patterns. In *Computer Graphics Forum*, volume 35, pages 281–290. Wiley Online Library.

Medeiros, D., Dos Anjos, R., Pantidi, N., Huang, K., Sousa, M., Anslow, C., and Jorge, J. (2021). Promoting reality awareness in virtual reality through proxemics. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 21–30. IEEE.

Mehmood, E. and Anees, T. (2020). Challenges and solutions for processing real-time big data stream: a systematic literature review. *IEEE Access*, 8:119123–119143.

Mehta, V. (2020). The new proxemics: Covid-19, social distancing, and sociable space. *Journal of Urban Design*, 25(6):669–674.

Mejova, Y., Srinivasan, P., and Boynton, B. (2013). Gop primary season on twitter: " popular" political sentiment in social media. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 517–526.

Mello, R. d. S., Bogorny, V., Alvares, L. O., Santana, L. H. Z., Ferrero, C. A., Frozza, A. A., Schreiner, G. A., and Renso, C. (2019). Master: A multiple aspect view on trajectories. *Transactions in GIS*, 23(4):805–822.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8.

Miles, A. and Bechhofer, S. (2009). Skos simple knowledge organization system reference. *W3C recommendation*.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. sage.

Miles, M. B. and Huberman, A. M. (2003). *Analyse des données qualitatives*. De Boeck Supérieur.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Misirlis, N. and Vlachopoulou, M. (2018). Social media metrics and analytics in marketing–s3m: A mapping literature review. *International Journal of Information Management*, 38(1):270–276.

Moens, M.-F., Li, J., and Chua, T.-S. (2014). *Mining user generated content*. CRC press.

Mohammad, S. M. and Turney, P. D. (2013). Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Mojan, J. and Sébastien, B. (2023). Phi-2: The surprising power of small language models — microsoft.com. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/. [Accessed 03-03-2024].

Moncla, L. and Gaio, M. (2023). Perdido: librairie python pour le geoparsing et le geocoding de textes en français. In *Extraction et Gestion des Connaissances (EGC'2023)*.

Moradi, M., Blagec, K., Haberl, F., and Samwald, M. (2021). Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555*.

Moreau, C., Devogele, T., Peralta, V., and Etienne, L. (2020). A contextual edit distance for semantic trajectories. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 635–637.

Morgan, M. B. H. and Van Keulen, M. (2014). Information extraction for social media. In *Proceedings of the Third Workshop on Semantic Web and Information Extraction*, pages 9–16.

Moro, S., Cortez, P., and Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent dirichlet allocation. *Expert Systems with Applications*, 42(3):1314–1324.

Mueller, F., Stellmach, S., Greenberg, S., Dippon, A., Boll, S., Garner, J., Khot, R., Naseem, A., and Altimira, D. (2014). Proxemics play: Understanding proxemics for designing digital play experiences. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, page 533–542, New York, NY, USA. Association for Computing Machinery.

Mulfari, D., Celesti, A., Fazio, M., Villari, M., and Puliafito, A. (2016). Using google cloud vision in assistive technology scenarios. In *2016 IEEE symposium on computers and communication (ISCC)*, pages 214–219. IEEE.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Murthy, K., Amminedu, E., and Rao, V. V. (2003). Integration of thematic maps through gis for identification of groundwater potential zones. *Journal of the Indian Society of Remote Sensing*, 31:197–210.

Murzintcev, N. and Cheng, C. (2017). Disaster hashtags in social media. *ISPRS International Journal of Geo-Information*, 6(7):204.

Mutanga, O. and Kumar, L. (2019). Google earth engine applications.

Naab, T. K. and Sehl, A. (2017). Studies of user-generated content: A systematic review. *Journalism*, 18(10):1256–1273.

Nadeau, D. (2007). *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision*. PhD thesis, University of Ottawa (Canada).

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California. Association for Computational Linguistics.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Nazir, A., Rao, Y., Wu, L., and Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Negash, S. (2004). Business intelligence. *Communications of the association for information systems*, 13(1):15.

Neiger, B. L., Thackeray, R., Van Wagenen, S. A., Hanson, C. L., West, J. H., Barnes, M. D., and Fagen, M. C. (2012). Use of social media in health promotion: purposes, key performance indicators, and evaluation metrics. *Health promotion practice*, 13(2):159–164.

Nemes, L. and Kiss, A. (2021). Social media sentiment analysis based on covid-19. *Journal of Information and Telecommunication*, 5(1):1–15.

Neteler, M. and Mitasova, H. (2002). *Open source GIS: a GRASS GIS approach*, volume 689. Springer Science & Business Media.

Ng, B. L., Liu, W., and Wang, J. C. (2016). Student motivation and learning in mathematics and science: A cluster analysis. *International Journal of Science and Mathematics Education*, 14:1359–1376.

Nguyen, T. T., Camacho, D., and Jung, J. E. (2017). Identifying and ranking cultural heritage resources on geotagged social media for smart cultural tourism services. *Personal and Ubiquitous Computing*, 21:267–279.

Nielsen, F. A. (2017). Afinn project. *DTU Compute Technical University of Denmark*.

Nouvel, D. (2012). Named entity recognition by mining association rules. *Theses, Université François Rabelais-Tours*.

Oba, A., Paik, I., and Kuwana, A. (2021). Automatic classification for ontology generation by pretrained language model. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 210–221. Springer.

OECD (2023). Indicateurs clés du tourisme | statistiques de l'ocde sur le tourisme. Accessed: 2023-11-20.

Ohana, B. and Tierney, B. (2009). Sentiment classification of reviews using sentiwordnet.

Oliveira, F. B., Mougouei, D., Haque, A., Sichman, J. S., Dam, H. K., Evans, S., Ghose, A., and Singh, M. P. (2023). Beyond fear and anger: A global analysis of emotional response to covid-19 news on twitter. *Online Social Networks and Media*, 36:100253.

O'reilly, T. (2009). *What is web 2.0.* " O'Reilly Media, Inc.".

Orlovskyi, D. and Kopp, A. (2020). A business intelligence dashboard design approach to improve data analytics and decision making. In Snytyuk, V., Anisimov, A., Krak, I., Nikitchenko, M., Marchenko, O., Mallet, F., Tsyganok, V. V., Aldrich, C., Pester, A., Tanaka, H., Henke, K., Chertov, O., Bozóki, S., and Vovk, V., editors, *Selected Papers of the 7th International Conference "Information Technology and Interactions" (IT&I-2020). Conferece Proceedings, Kyiv, Ukraine, December 02-03, 2020*, volume 2833 of *CEUR Workshop Proceedings*, pages 48–59. CEUR-WS.org.

Owuor, I. and Hochmair, H. H. (2020). An overview of social media apps and their potential role in geospatial research. *ISPRS international journal of geo-information*, 9(9):526.

Pablo-Romero, M. d. P., Pozo-Barajas, R., and Sánchez-Rivas, J. (2019). Tourism and temperature effects on the electricity consumption of the hospitality sector. *Journal of cleaner production*, 240:118168.

Pan, G. and Pan, J. (2012). Research in crop land suitability analysis based on gis. In *Computer and Computing Technologies in Agriculture V: 5th IFIP TC 5/SIG 5.1 Conference, CCTA 2011, Beijing, China, October 29-31, 2011, Proceedings, Part II 5*, pages 314–325. Springer.

Paolanti, M., Mancini, A., Frontoni, E., Felicetti, A., Marinelli, L., Marcheggiani, E., and Pierdicca, R. (2021). Tourism destination management using sentiment analysis and geo-location information: a deep learning approach. *Information Technology & Tourism*, 23(2):241–264.

Parekh, D., Amarasingam, A., Dawson, L., and Ruths, D. (2018). Studying jihadists on social media: A critique of data collection methodologies. *Perspectives on Terrorism*, 12(3):5–23.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Park, Y. and Jo, I.-H. (2015). Development of the learning analytics dashboard to support students' learning performance. *Journal of Universal Computer Science*, 21(1):110.

Parsons, A. (2013). Using social media to reach consumers: A content analysis of official facebook pages. *Academy of marketing studies Journal*, 17(2):27.

Penn, G. and Carpendale, S. (2009). Linguistic information visualization. https://esslli2009.labri.fr/course_82.html. ESSLLI 2009, Language and Computation advanced course.

Pérez, P., Roose, P., Cardinale, Y., Dalmau, M., Masson, D., and Couture, N. (2021). An approach to develop mobile proxemic applications. *J. Data Intell.*, 2(2):166–189.

Piedrahita-Valdés, H., Castillo, D., Bermejo, J., Guillem-Saiz, P., Higuera, J.-R., Guillem-Saiz, J., Montalvo, J. A., and Machio, F. (2021). Vaccine hesitancy on social media: Sentiment analysis from june 2011 to april 2019. *Vaccines*, 9:28.

Pike, S. and Page, S. J. (2014). Destination marketing organizations and destination marketing: A narrative analysis of the literature. *Tourism management*, 41:202–227.

Pilat Tourisme (2022). Tableau de bord 2022. Accessed: 2023-11-20.

Priambodo, R. and Satria, R. (2012). User behavior pattern of mobile online social network service. In *2012 International Conference on Cloud Computing and Social Networking (ICCCSN)*, pages 1–4. IEEE.

Punchoojit, L., Hongwarittorrn, N., et al. (2017). Usability studies on mobile user interface design patterns: a systematic literature review. *Advances in Human-Computer Interaction*, 2017.

Pérez, J. M., Giudici, J. C., and Luque, F. (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Quiña-Mera, A., Fernandez, P., García, J. M., and Ruiz-Cortés, A. (2023). Graphql: a systematic mapping study. *ACM Computing Surveys*, 55(10):1–35.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Raghavendra, S. (2021). Introduction to selenium. *Python Testing with Selenium: Learn to Implement Different Testing Techniques Using the Selenium WebDriver*, pages 1–14.

Rahman, Z. et al. (2017). The impact of social media engagement metrics on purchase intention: A study on brand fan page followers. *LUMEN Proceedings*, 1:665–681.

Rajaonarivo, L., Mine, T., and Arakawa, Y. (2022). Coupling of semantic and syntactic graphs generated via tweets to detect local events. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 128–133. IEEE.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Rathore, A. K., Kar, A. K., and Ilavarasan, P. V. (2017). Social media analytics: Literature review and directions for future research. *Decision Analysis*, 14(4):229–249.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155.

Rehurek, R. and Sojka, P. (2011). Gensim—statistical semantics in python. *Retrieved from genism. org*.

Reinhart, C. F. and Wienold, J. (2011). The daylighting dashboard–a simulation-based design analysis for daylit spaces. *Building and environment*, 46(2):386–396.

Reyes, A., Rosso, P., and Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Richardson, L. (2007). Beautiful soup documentation.

Rios-Martinez, J., Spalanzani, A., and Laugier, C. (2015). From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, 7(2):137–153.

Roberts, L. D., Howell, J. A., and Seaman, K. (2017). Give me a customizable dashboard: Personalized learning analytics dashboards in higher education. *Technology, Knowledge and Learning*, 22:317–333.

Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., and Gribov, A. (2018). Rusentiment: An enriched sentiment analysis dataset for social media in russian. In *Proceedings of the 27th international conference on computational linguistics*, pages 755–763.

Rogers, D., Preece, A., Innes, M., and Spasić, I. (2021). Real-time text classification of user-generated content on social media: Systematic review. *IEEE Transactions on Computational Social Systems*, 9(4):1154–1166.

Rose, J., Mackey-Kallis, S., Shyles, L., Barry, K., Biagini, D., Hart, C., and Jack, L. (2012). Face it: The impact of gender on social media images. *Communication Quarterly*, 60(5):588–607.

Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado. Association for Computational Linguistics.

Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland. Association for Computational Linguistics.

Roser, M., Ritchie, H., and Ortiz-Ospina, E. (2015). Internet. *Our World in Data*. https://ourworldindata.org/internet#the-rise-of-social-media.

Rout, J. K., Choo, K.-K. R., Dash, A. K., Bakshi, S., Jena, S. K., and Williams, K. L. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, 18:181–199.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121.

Ruch, P., Baud, R., Geissbühler, A., and Rassinoux, A.-M. (2001). Comparing general and medical texts for information retrieval based on natural language processing: an inquiry into lexical disambiguation. In *MEDINFO 2001*, pages 261–265. IOS Press.

Sadilek, A., Kautz, H., and Silenzio, V. (2012). Modeling spread of disease from social interactions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 322–329.

Sadri, A. M., Hasan, S., Ukkusuri, S. V., and Suarez Lopez, J. E. (2018). Analysis of social interaction network properties and growth on twitter. *Social Network Analysis and Mining*, 8:1–13.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.

Saif, H., He, Y., and Alani, H. (2012). Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11*, pages 508–524. Springer.

Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O. L., Rigau, G., and Agirre, E. (2024). GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Sanh, V. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Santia, G. and Williams, J. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the international AAAI conference on web and social media*, volume 12, pages 531–540.

Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, 2(5):377.

Sathick, J. and Venkat, J. (2015). A generic framework for extraction of knowledge from social web sources (social networking websites) for an online recommendation system. *International Review of Research in Open and Distributed Learning*, 16(2):247–271.

Savova, G. K., Coden, A. R., Sominsky, I. L., Johnson, R., Ogren, P. V., De Groen, P. C., and Chute, C. G. (2008). Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of biomedical informatics*, 41(6):1088–1100.

Scepanovic, S., Martin-Lopez, E., Quercia, D., and Baykaner, K. (2020). Extracting medical entities from social media. In *Proceedings of the ACM conference on health, inference, and learning*, pages 170–181.

Scheibel, W., Trapp, M., Limberger, D., and Döllner, J. (2020). A taxonomy of treemap visualization techniques. In *VISIGRAPP (3: IVAPP)*, pages 273–280.

Schick, T. and Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Scholz, J. and Jeznik, J. (2020). Evaluating geo-tagged twitter data to analyze tourist flows in styria, austria. *ISPRS International Journal of Geo-Information*, 9(11):681.

Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., and Mühlhäuser, M. (2013). A multi-indicator approach for geolocalization of tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 573–582.

Scott, D., Gössling, S., and de Freitas, C. R. (2008). Preferred climates for tourism: case studies from canada, new zealand and sweden. *Climate Research*, 38(1):61–73.

Seethal (2023). Sentiment analysis generic dataset. https://huggingface.co/Seethal/sentiment_analysis_generic_dataset. Accessed on 23th March 2023.

Serna, A., Soroa, A., and Agerri, R. (2021). Applying deep learning techniques for sentiment analysis to assess sustainable transport. *Sustainability*, 13(4).

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., and Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18–38.

Shannon Greenwood, A. P. and Duggan, M. (2016). Social Media Update 2016 — pewresearch.org. [Accessed 27-03-2024].

Sharma, S. S. and Dutta, G. (2021). Sentidraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination. *Information Processing & Management*, 58(1):102412.

Sherly, S., Halim, F., and Sudirman, A. (2020). The role of social media in increasing market share of msme products in pematangsiantar city. *Jurnal Manajemen Dan Bisnis*, 9(2):61–72.

Shimada, K., Inoue, S., Maeda, H., and Endo, T. (2011). Analyzing tourism information on twitter for a local city. In *2011 First ACIS International Symposium on Software and Network Engineering*, pages 61–66. IEEE.

Shinan, K., Alsubhi, K., and Ashraf, M. U. (2023). Botsward: Centrality measures for graph-based bot detection using machine learning. *Computers, Materials & Continua*, 75(1).

Shneiderman, B. (2004). Designing for fun: how can we design user interfaces to be more fun? *interactions*, 11(5):48–50.

Shukla, A. and Dhir, S. (2016). Tools for data visualization in business intelligence: case study using the tool qlikview. In *Information Systems Design and Intelligent Applications: Proceedings of Third International Conference INDIA 2016, Volume 2*, pages 319–326. Springer.

Siabato, W., Claramunt, C., Ilarri, S., and Manso-Callejo, M. Á. (2018). A survey of modelling trends in temporal gis. *ACM Computing Surveys (CSUR)*, 51(2):1–41.

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Slavkovikj, V., Verstockt, S., Van Hoecke, S., and Van de Walle, R. (2014). Review of wildfire detection using social media. *Fire safety journal*, 68:109–118.

Sloan, L. and Morgan, J. (2015). Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PloS one*, 10(11):e0142209.

Słomska-Przech, K. and Gołębiowska, I. M. (2021). Do different map types support map reading equally? comparing choropleth, graduated symbols, and isoline maps for map use tasks. *ISPRS International Journal of Geo-Information*, 10(2):69.

Smailhodzic, E., Hooijsma, W., Boonstra, A., and Langley, D. J. (2016). Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. *BMC health services research*, 16(1):1–14.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013a). Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Souili, A., Cavallucci, D., and Rousselot, F. (2015). Natural language processing (nlp)–a solution for knowledge extraction from patent unstructured data. *Procedia engineering*, 131:635–643.

Sponcil, M. and Gitimu, P. (2013). Use of social media by college students: Relationship to communication and self-concept. *Journal of Technology Research*, 4(1):37–49.

Stadler, J. G., Donlon, K., Siewert, J. D., Franken, T., and Lewis, N. E. (2016). Improving the efficiency and ease of healthcare analysis through use of data visualization dashboards. *Big data*, 4(2):129–135.

Statista (2024). User-generated internet content per minute 2023 | Statista — statista.com. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/. [Accessed 30-03-2024].

Stieglitz, S., Mirbabaie, M., Ross, B., and Neuberger, C. (2018). Social media analytics–challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39:156–168.

Suarez-Lledo, V. and Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, 23(1):e17187.

Sui, D. and Goodchild, M. (2011). The convergence of gis and social media: challenges for giscience. *International journal of geographical information science*, 25(11):1737–1748.

Sun, P., Yang, X., Zhao, X., and Wang, Z. (2018). An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278. IEEE.

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Suzuki, S. (2020). Use of online travel agencies as a data source for tourism marketing. *Journal of Global Tourism Research*, 5(2):167–171.

Szafir, D. A. (2017). Modeling color difference for visualization design. *IEEE transactions on visualization and computer graphics*, 24(1):392–401.

T. K. Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

T. K. Sang, E. F. and Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Tahri, M., Hakdaoui, M., and Maanan, M. (2015). The evaluation of solar farm locations applying geographic information system and multi-criteria decision-making methods: Case study in southern morocco. *Renewable and sustainable energy reviews*, 51:1354–1362.

Tang, J., Chang, Y., Aggarwal, C., and Liu, H. (2016). A survey of signed network mining in social media. *ACM Computing Surveys (CSUR)*, 49(3):1–37.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Terryn, A. R., Drouin, P., Hoste, V., and Lefever, E. (2019). Analysing the impact of supervised machine learning on automatic term extraction: Hamlet vs termostat. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1012–1021.

Theiss, S. K., Burke, R. M., Cory, J. L., and Fairley, T. L. (2016). Getting beyond impressions: an evaluation of engagement with breast cancer-related facebook content. *Mhealth*, 2.

Thoenig, J.-C. (2010). Politique publique. *Dictionnaire des politiques publiques*, pages 420–427.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tonia, T., Van Oyen, H., Berger, A., Schindler, C., and Künzli, N. (2016). If i tweet will you cite? the effect of social media exposure of articles on downloads and citations. *International journal of public health*, 61:513–520.

Toporkov, O. and Agerri, R. (2024). On the role of morphological information for contextual lemmatization. *Computational Linguistics*, pages 1–35.

Tosi, S. (2009). *Matplotlib for Python developers*. Packt Publishing Ltd.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tripathy, J. K., Chakkaravarthy, S. S., Satapathy, S. C., Sahoo, M., and Vaidehi, V. (2022). Albert-based fine-tuning model for cyberbullying analysis. *Multimedia Systems*, 28(6):1941–1949.

Trunfio, M. and Rossi, S. (2021). Conceptualising and measuring social media engagement: A systematic literature review. *Italian Journal of Marketing*, 2021(3):267–292.

Tsao, S.-F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., and Butt, Z. A. (2021). What social media told us in the time of covid-19: a scoping review. *The Lancet Digital Health*, 3(3):e175–e194.

Tsou, M.-H. and Curran, J. M. (2008). User-centered design approaches for web mapping applications: A case study with usgs hydrological data in the united states. In *International perspectives on maps and the Internet*, pages 301–321. Springer.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., and Pereg, O. (2022). Efficient few-shot learning without prompts.

Turcan, E. and Mckeown, K. (2019). Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107.

UNWTO (2023). Tableau de bord de l'omt de données sur le tourisme. Accessed: 2023-11-20.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4(1):14–28.

Van Bruwaene, D., Huang, Q., and Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54:851–874.

Van Eck, N. J. and Waltman, L. (2013). Vosviewer manual. *Leiden: Univeristeit Leiden*, 1(1):1–53.

Vannucci, A., Simpson, E. G., Gagnon, S., and Ohannessian, C. M. (2020). Social media use and risky behaviors in adolescents: A meta-analysis. *Journal of adolescence*, 79:258–274.

Varlamis, I., Sardianos, C., Bogorny, V., Alvares, L. O., Carvalho, J. T., Renso, C., Perego, R., and Violos, J. (2021). A novel similarity measure for multiple aspect trajectory clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 551–558.

Vashisht, V. and Dharia, P. (2020). Integrating chatbot application with qlik sense business intelligence (bi) tool using natural language processing (nlp). In *Micro-Electronics and Telecommunication Engineering: Proceedings of 3rd ICMETE 2019*, pages 683–692. Springer.

Vásquez, J., Gómez-Adorno, H., and Bel-Enguix, G. (2021). Bert-based approach for sentiment analysis of spanish reviews from tripadvisor. In *IberLEF@ SEPLN*, pages 165–170.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vázquez-Ingelmo, A., Garcia-Penalvo, F. J., and Theron, R. (2019). Information dashboards and tailoring capabilities-a systematic literature review. *IEEE Access*, 7:109673–109688.

Venegas-Vera, A. V., Colbert, G. B., and Lerma, E. V. (2020). Positive and negative impact of social media in the covid-19 era. *Reviews in cardiovascular medicine*, 21(4):561–564.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., and Pustejovsky, J. (2009). The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43:161–179.

Viñán-Ludeña, M. S. and de Campos, L. M. (2021). Analyzing tourist data on twitter: a case study in the province of granada at spain. *Journal of Hospitality and Tourism Insights*.

Visit Paris Region (2023). Tableau de bord. Accessed: 2023-11-20.

Walker, O., Simpson, G. D., Teo, A. C., and Newsome, D. (2019). Preprint: Crowdsourcing and analysing wildlife tourism data from photographs shared on social media. *Preprints*.

Wang, K. and Chua, T.-S. (2010). Exploiting salient patterns for question detection and question retrieval in community-based question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1155–1163.

Wang, L., Li, R., Yan, Y., Yan, Y., Wang, S., Wu, W., and Xu, W. (2022). Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint*, 2203.03903.

Wang, L., Ramachandran, A., and Chaintreau, A. (2016a). Measuring click and share dynamics on social media: a reproducible and validated approach. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 108–113.

Wang, W. (2012). Chinese news event 5w1h semantic elements extraction for event ontology population. In *Proceedings of the 21st International Conference on World Wide Web*, pages 197–202.

Wang, W., Zhang, G., and Lu, J. (2016b). Member contribution-based group recommender system. *Decision Support Systems*, 87:80–93.

Wang, W., Zhao, D., Zou, L., Wang, D., and Zheng, W. (2010a). Extracting 5w1h event semantic elements from chinese online news. In *Web-Age Information Management: 11th International Conference, WAIM 2010, Jiuzhaigou, China, July 15-17, 2010. Proceedings 11*, pages 644–655. Springer.

Wang, X., Tang, L., Gao, H., and Liu, H. (2010b). Discovering overlapping groups in social media. In *2010 IEEE international conference on data mining*, pages 569–578. IEEE.

Welsh, M. E. (2014). Review of voyant tools. *Collaborative Librarianship*, 6(2):96–98.

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., et al. (2017). The comptox chemistry dashboard: a community data resource for environmental chemistry. *Journal of cheminformatics*, 9:1–27.

Williamson, J., Li, J., Vinayagamoorthy, V., Shamma, D. A., and Cesar, P. (2021). Proxemics and social interactions in an instrumented virtual reality workshop. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218.

Wittwer, M., Reinhold, O., Alt, R., Jessen, F., and Stüber, R. (2017). Social media analytics using business intelligence and social media tools–differences and implications. In *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers 19*, pages 252–259. Springer.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

World Tourism Organization (2002). *Thesaurus on Tourism and Leisure Activities*. World Tourism Organization.

Wright, E., Khanfar, N. M., Harrington, C., Kizer, L. E., et al. (2010). The lasting effects of social media trends on advertising. *Journal of Business & Economics Research (JBER)*, 8(11).

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics, 1994*, pages 133–138.

Xia, F., Liu, J., Nie, H., Fu, Y., Wan, L., and Kong, X. (2019). Random walks: A review of algorithms and applications. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2):95–107.

Xu, L., Dong, Q., Liao, Y., Yu, C., Tian, Y., Liu, W., Li, L., Liu, C., Zhang, X., et al. (2020). Cluener2020: fine-grained named entity recognition dataset and benchmark for chinese. *arXiv preprint arXiv:2001.04351*.

Yadav, H. and Sagar, M. (2023). Exploring covid-19 vaccine hesitancy and behavioral themes using social media big-data: a text mining approach. *Kybernetes*, 52(7):2616–2648.

Yang, S., Jiang, X., Zhao, H., Zeng, W., Liu, H., and Jia, Y. (2024). Faima: Feature-aware in-context learning for multi-domain aspect-based sentiment analysis. *arXiv preprint arXiv:2403.01063*.

Yang, Y., Baker, S., Kannan, A., and Ramanan, D. (2012). Recognizing proxemics in personal photos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3522–3529.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yeginbergen, A. and Agerri, R. (2024). Cross-lingual Argument Mining in the Medical Domain. *arXiv 2301.10527*.

Yeh, A., Ratsamee, P., Kiyokawa, K., Uranishi, Y., Mashita, T., Takemura, H., Fjeld, M., and Obaid, M. (2017). Exploring proxemics for human-drone interaction. In *Proceedings of the 5th International*

Conference on Human Agent Interaction, HAI '17, page 81–88, New York, NY, USA. Association for Computing Machinery.

Yohanna, A. (2020). The influence of social media on social interactions among students. *Indonesian Journal of Social Sciences*, 12(2):34–48.

Yoo, E., Rabinovich, E., and Gu, B. (2020). The growth of follower networks on social media platforms for humanitarian operations. *Production and Operations Management*, 29(12):2696–2715.

Yu, X. and Yuan, C. (2019). How consumers' brand experience in social media can improve brand perception and customer equity. *Asia Pacific Journal of Marketing and Logistics*, 31(5):1233–1251.

Yue, L., Chen, W., Li, X., Zuo, W., and Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60:617–663.

Zang, T., Zhu, Y., Liu, H., Zhang, R., and Yu, J. (2022). A survey on cross-domain recommendation: taxonomies, methods, and future directions. *ACM Transactions on Information Systems*, 41(2):1–39.

Zangerle, E., Gassler, W., and Specht, G. (2013). On the impact of text similarity functions on hashtag recommendations in microblogging environments. *Social network analysis and mining*, 3:889–898.

Zarei, K., Farahbakhsh, R., Crespi, N., and Tyson, G. (2020). A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226*.

Zhang, S. (2003). *Thematic knowledge extraction*. Nottingham Trent University (United Kingdom).

Zhang, Y., Wang, X., Sakai, Y., and Yamasaki, T. (2019). Measuring similarity between brands using followers' post in social media. *Proceedings of the ACM Multimedia Asia*, pages 1–6.

Zhang, Z. (2013). *Named entity recognition: challenges in document annotation, gazetteer construction and disambiguation*. PhD thesis, University of Sheffield.

Zhao, X., Sala, A., Wilson, C., Wang, X., Gaito, S., Zheng, H., and Zhao, B. Y. (2012). Multi-scale dynamics in a massive online social network. In *Proceedings of the 2012 Internet Measurement Conference*, pages 171–184.

Zheng, S., Wang, J., Sun, C., Zhang, X., and Kahn, M. E. (2019). Air pollution lowers chinese urbanites' expressed happiness on social media. *Nature human behaviour*, 3(3):237–243.

Zhu, F., Wang, Y., Chen, C., Zhou, J., Li, L., and Liu, G. (2021). Cross-domain recommendation: challenges, progress, and prospects. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 4721–4728. International Joint Conferences on Artificial Intelligence.

Zhu, N. Q. (2013). *Data visualization with D3. js cookbook*. Packt Publishing Ltd.

Zotova, E., Agerri, R., and Rigau, G. (2021). Semi-automatic generation of multilingual datasets for stance detection in twitter. *Expert Systems with Applications*, 170:114547.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., and Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.