REUNIÓN DE XXIV ANIVERSARIO
ASOCIACIÓN VENEZOLANA DE GESTIÓN DE INVESTIGACIÓN Y DESARROLLO AVEGID
ASOCIACIÓN INTERNACIONAL DE GESTIÓN DE INVESTIGACIÓN Y DESARROLLO AIGID

Optimal Strategies to Perform Multilingual Analysis of Social Content in the Tourism Domain

Maxime Masson, Christian Sallaberry, Rodrigo Agerri, Marie-Noelle Bessagnet, Philippe Roose, Annig Le Parc Lacayrelle

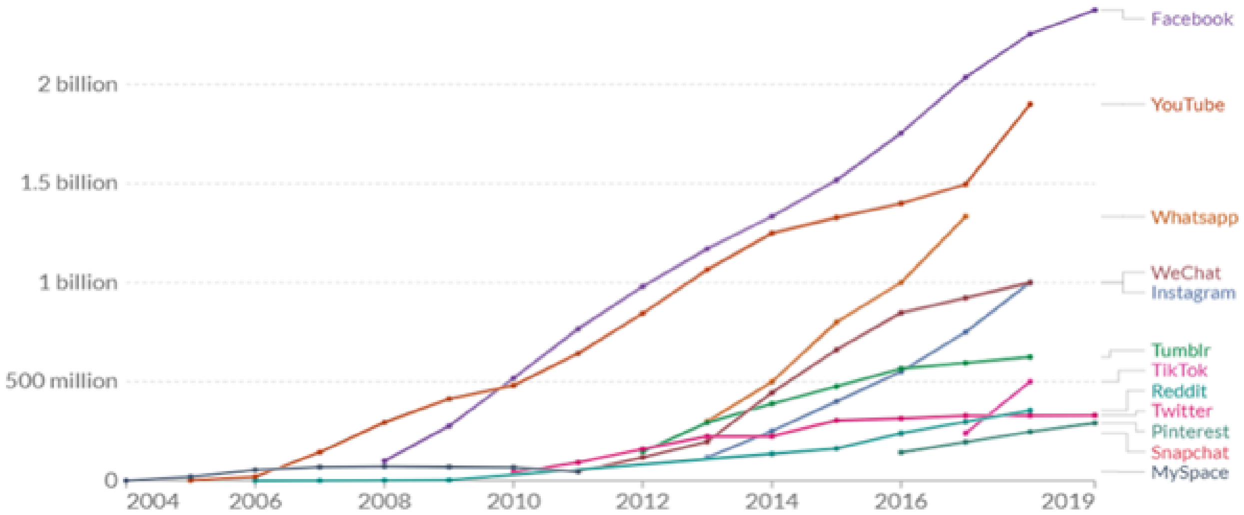
Caracas, 30 de enero de 2024

Context

User-Generated Content and Social Media

Significant growth in **data sources** available in many domains

Web 2.0 and **User-Generated Content (UGC)**



Source: *Our World In Data*

48.3% of the **world's population** using **social media** in 2020

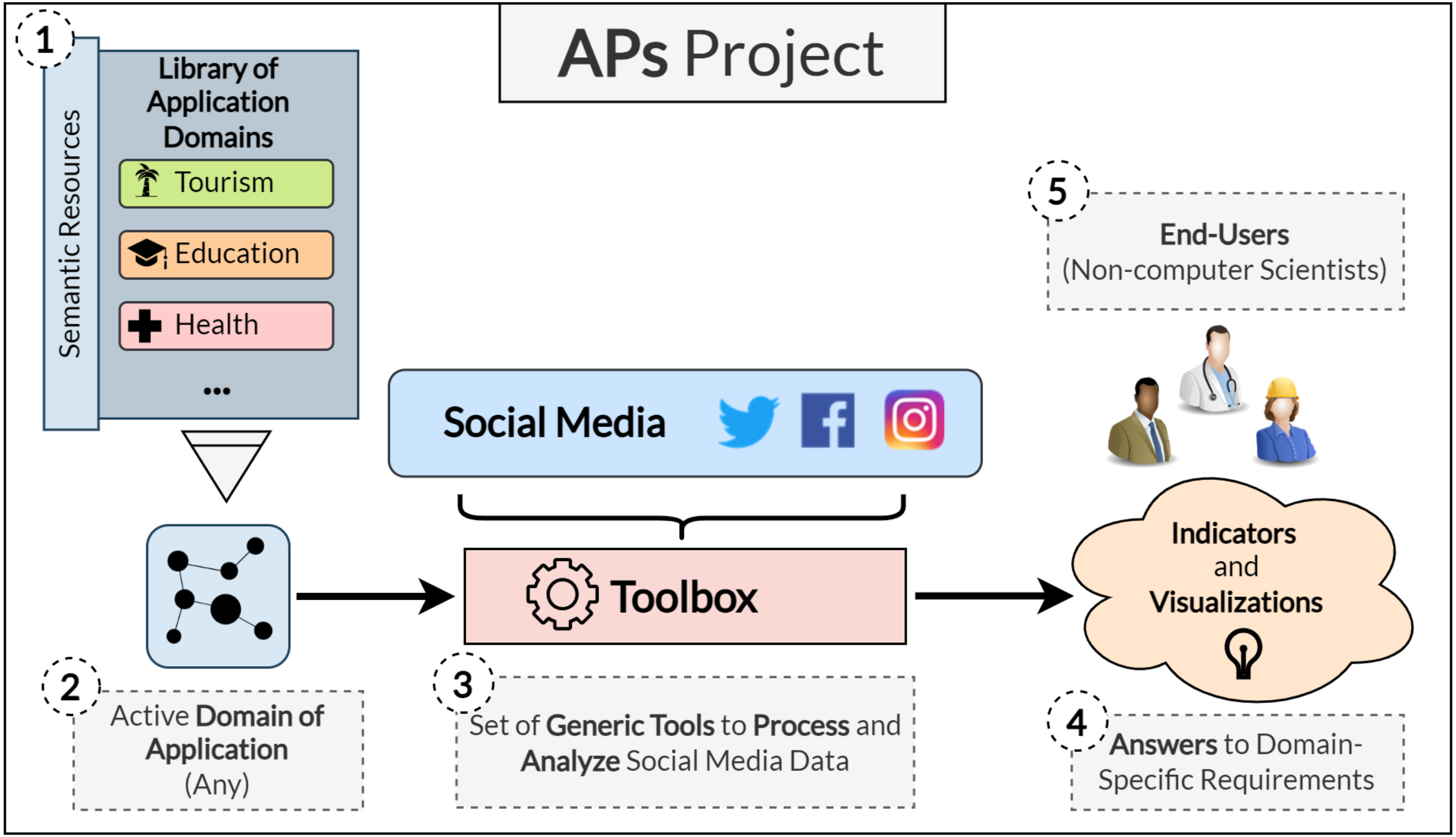
Massive content (per day)

- 500 millions *X* tweets
- 216 millions *Facebook* messages

Source: *Statista*

Context

APs Project



Motivations

Decision Support in the Tourism Domain

Tourism Professionals

- Assistance in the **decision-making** process and infrastructure planning
- Understanding the **requirements**, **practices** and **expectations** of visitors

A. What leisure activities do tourists typically engage in together?

B. Which cities do tourists tend to go to after visiting Bayonne?

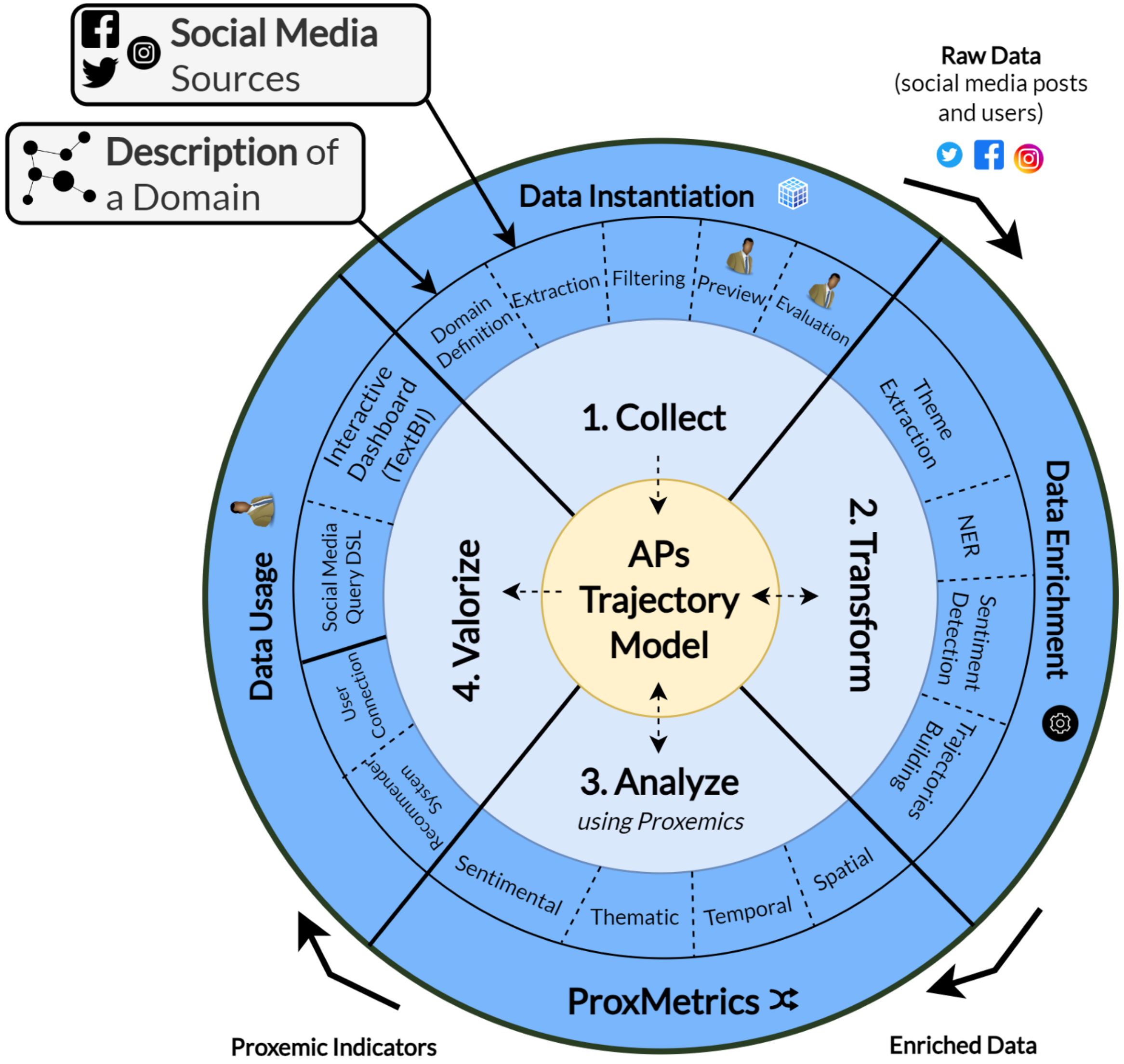
C. What are the typical demographics of tourists who visit Biarritz?

D. What are the common chained tourist activities?

Framework

Life Cycle

- ❑ **Double Genericity**
  - ❑ Domain of Application
  - ❑ Social Media Source
- ❑ **Semantic resource** to describe the domain
- ❑ **Decision support** for **non-technical users**
  - ❑ Indicators and Visualizations



Information Extraction

Common Knowledge Extraction Tasks

Sentiment Analysis

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

Named Entity Recognition (NER) for Locations

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

Fine-Grained Thematic Concept Extraction

Yesterday, we went swimming [Sports::WaterSports::Swimming] at the Grande Plage beach [NaturalResources:Beaches] in Biarritz, it was an amazing experience! The weather [NaturalResources::Weather] was incredibly sunny [ClimaticFactors::Sun]. :)

holidays [VisitorFlows::Holidays] # sun [ClimaticFactors::Sun] #btz

Research Challenges

Knowledge Extraction for Social Media Content

- **Challenge 1**: Identifying best knowledge extraction strategies and models for a given application domain and task
- **Challenge 2**: Determining the number of domain-specific annotated examples needed to get satisfying results
  - Manually annotating → **Lengthy, costly, time-consuming**
    - Objective → Get optimal results with **minimal use** of annotated examples
- **Hypothesis**: Use of a comparative study on optimal strategies for multilingual analysis of social media content based on a novel annotated dataset
- **Limitation**: Tourism domain, social media texts

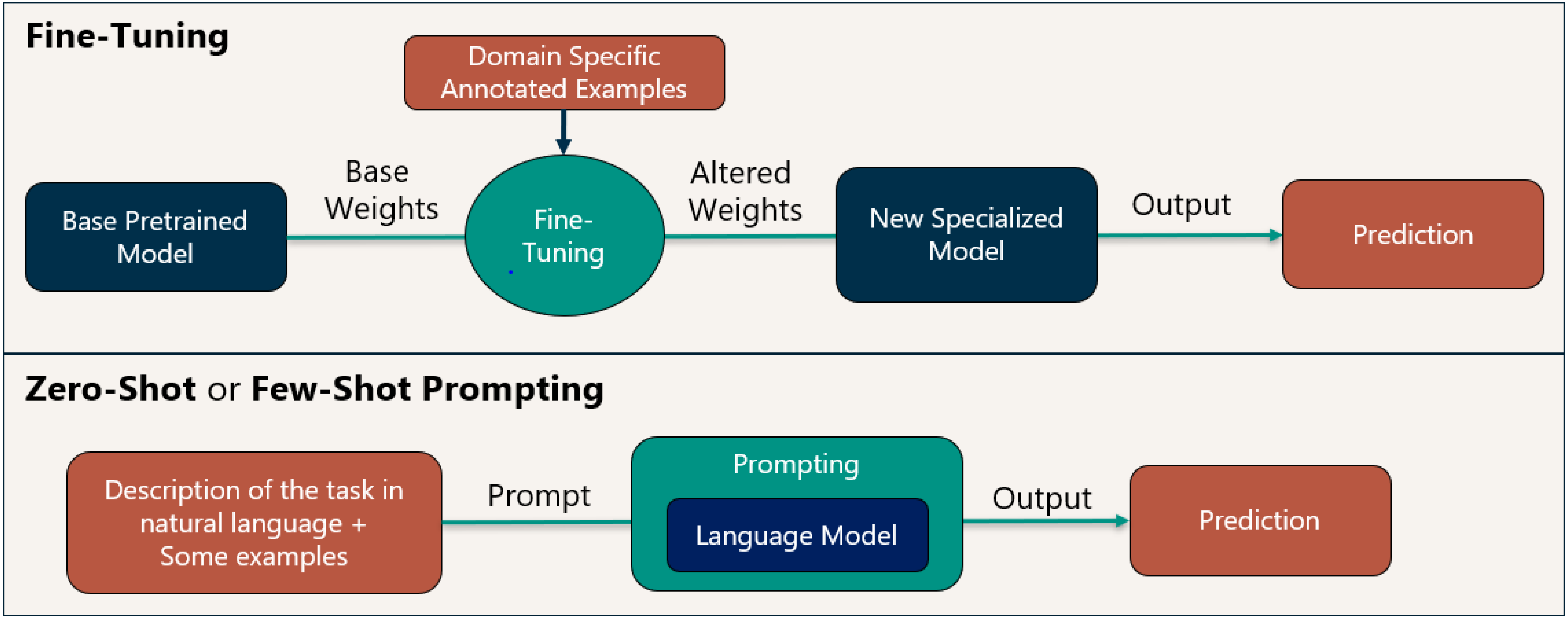
Related Work

Rules-based Techniques

Technique based on …	Advantages (+)	Disadvantages (-)
Lexicon	**Easy to implement,** easily understandable	Requires a lexicon, **limited by lexicon size**, **ignores context** and grammatical structures
Patterns	Precise for **well-defined** patterns	Missing **variations not covered by patterns**, requires **properly formatted sentences**
Syntax and grammar	Exploits **linguistic structures** for deeper analysis	**Complex to maintain**, especially in **multilingual contexts**
Semantics	Can understand **nuanced meanings** and **relationships between terms**	**Require comprehensive semantic knowledge** bases, more computationally intensive

Related Work

Fine-Tuning and Few-Shot Prompting

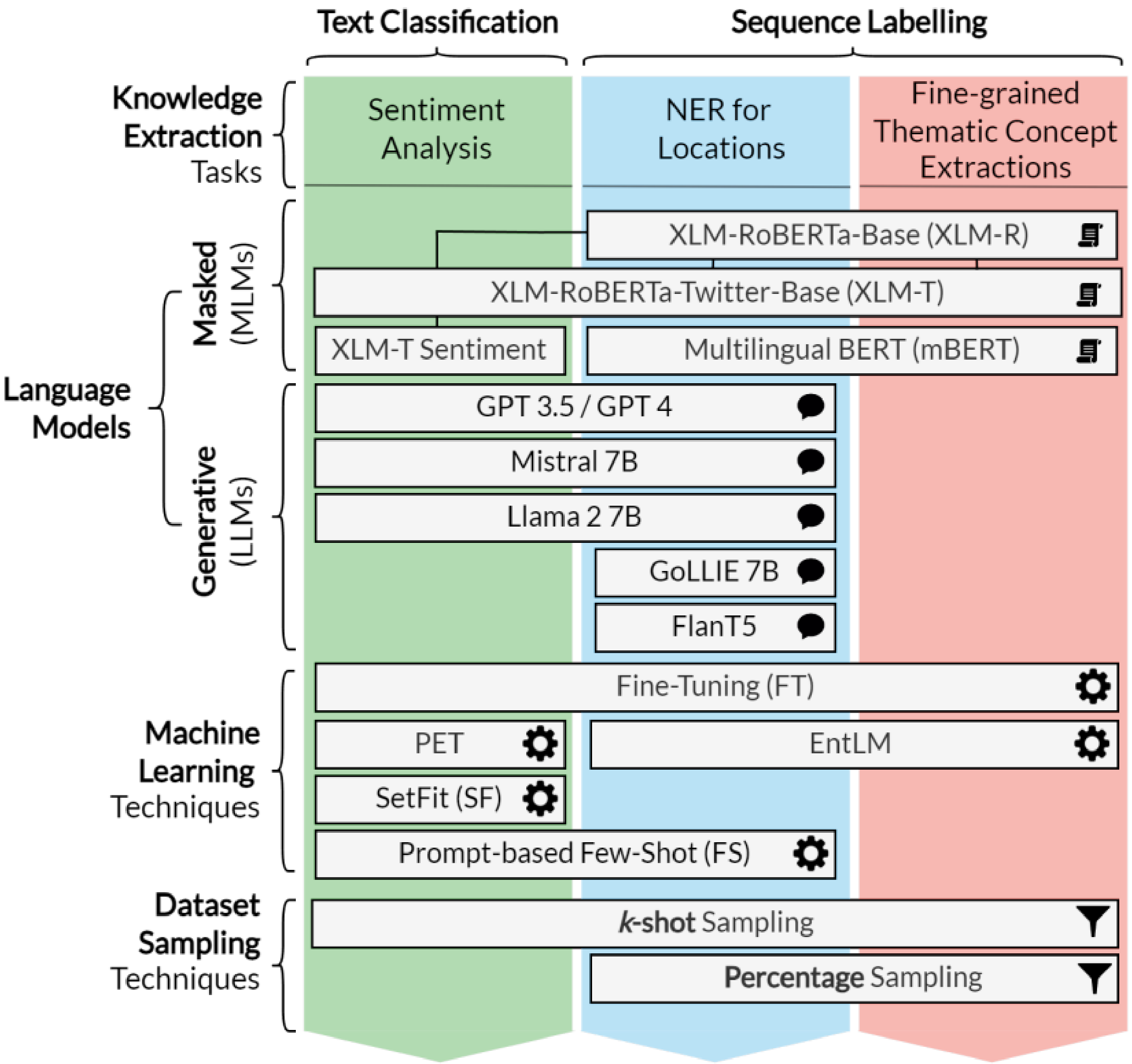


Experimental Setup

Comparative Analysis

- **Multilingual Dataset**
  - 2961 tweets, 624 users

1. Language Models
2. Machine Learning Techniques
3. Dataset Sampling Methods



Results

Sentiment Analysis

- ✓ A model (like XLM-T Sentiment) pre-trained on a **large dataset even out of domain** surpasses all other techniques
  - Best results are achieved with **just 5 examples**
    - Very high precision of 0.939
- ✓ If **such a model is not available**, few-shot prompting techniques **with LLMs** (like GPT and Mistral) are the **best alternatives** to avoid lengthy manual annotation.
  - Accuracy of 0.785 for GPT 3.5, 0.716 for Mistral 7B

	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

Results

NER for Locations

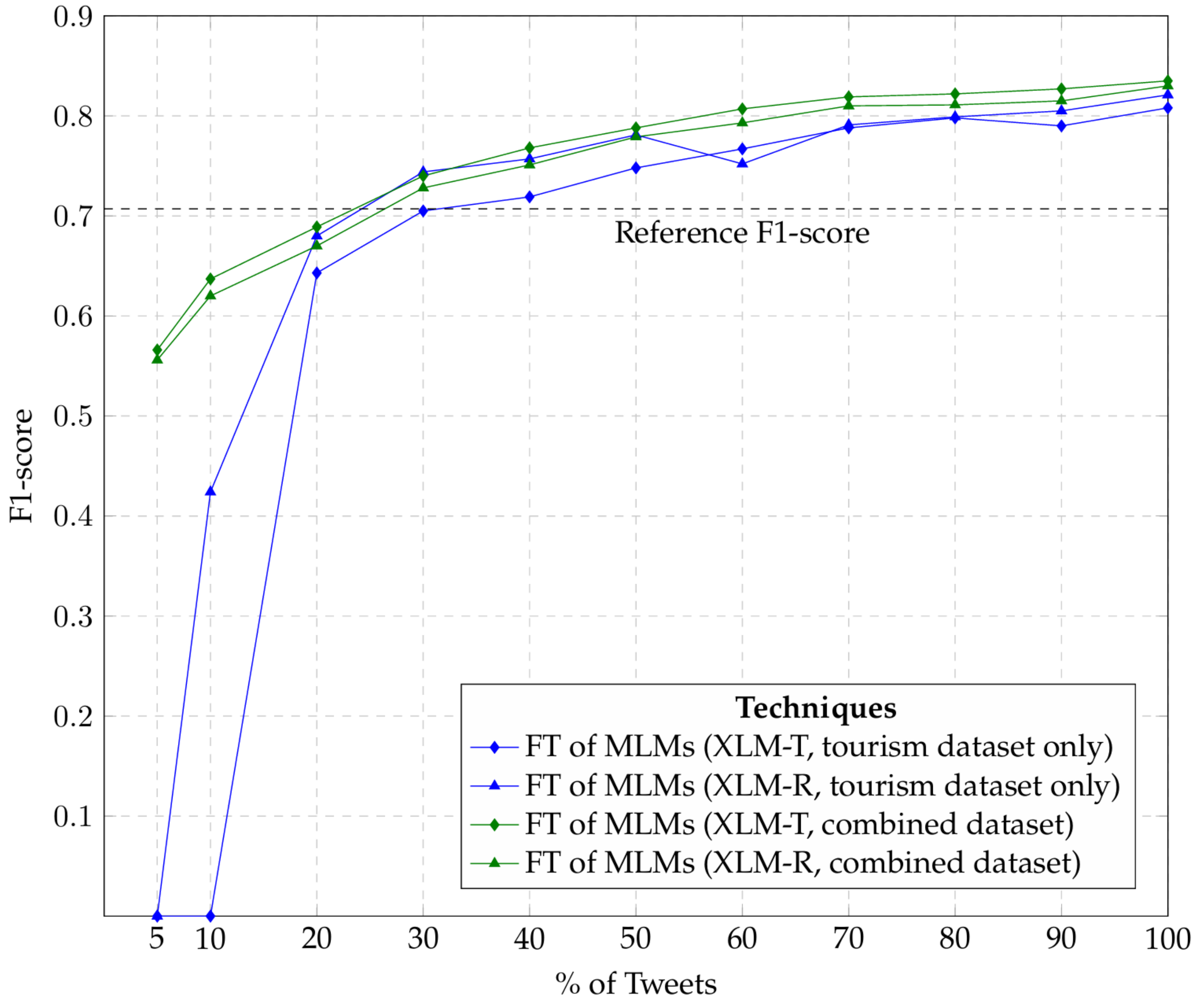
- Task with **few classes** (1 LOC class) but **many label words** (625 toponyms, **low representativeness of label words**)
- If it is possible **to annotate many examples** (330 minimum and up to 800 for optimal results), **fine-tuning with MLMs** works very well (F1 > 0.8 with 800 examples)
- Otherwise, **30 examples are enough** to **obtain satisfactory results** in few-shot with LLMs (0.75 with Mistral and 30 examples)
  - With **only 10 examples** , the machine learning-based methods **outperform the rule-based method**

	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**

Results

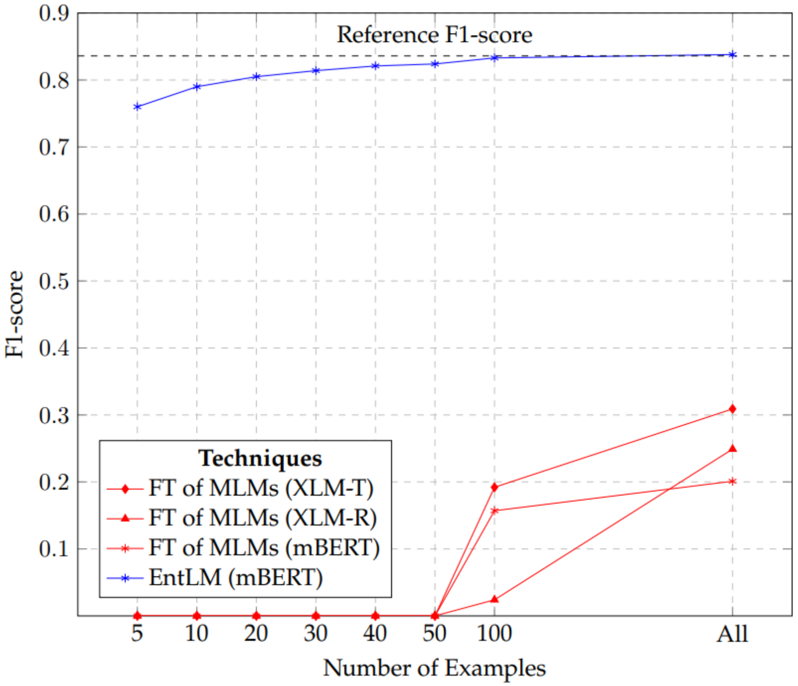
Discussion

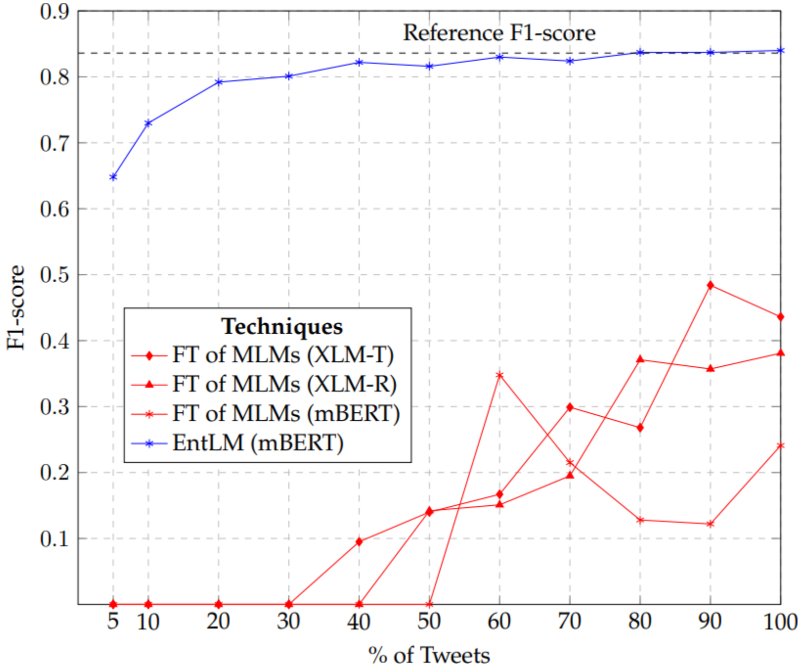
- ✓ **Combining** our training dataset with other datasets dedicated to NER
  - ESTER Corpus
  - AnCora
  - Broad Twitter Corpus (BTC)
- ✓ **No significant improvements**



Results

Fine-grained Thematic Concept Extraction





- ✓ Task with **many classes** (315 classes of the tourism thesaurus) but **few label words** (high representativeness of label words)
- ✓ **Transfer learning** (fine-tuning) does not work
- ✓ **EntLM** with ~1000 annotated tweets or lexicon-based approach is preferred

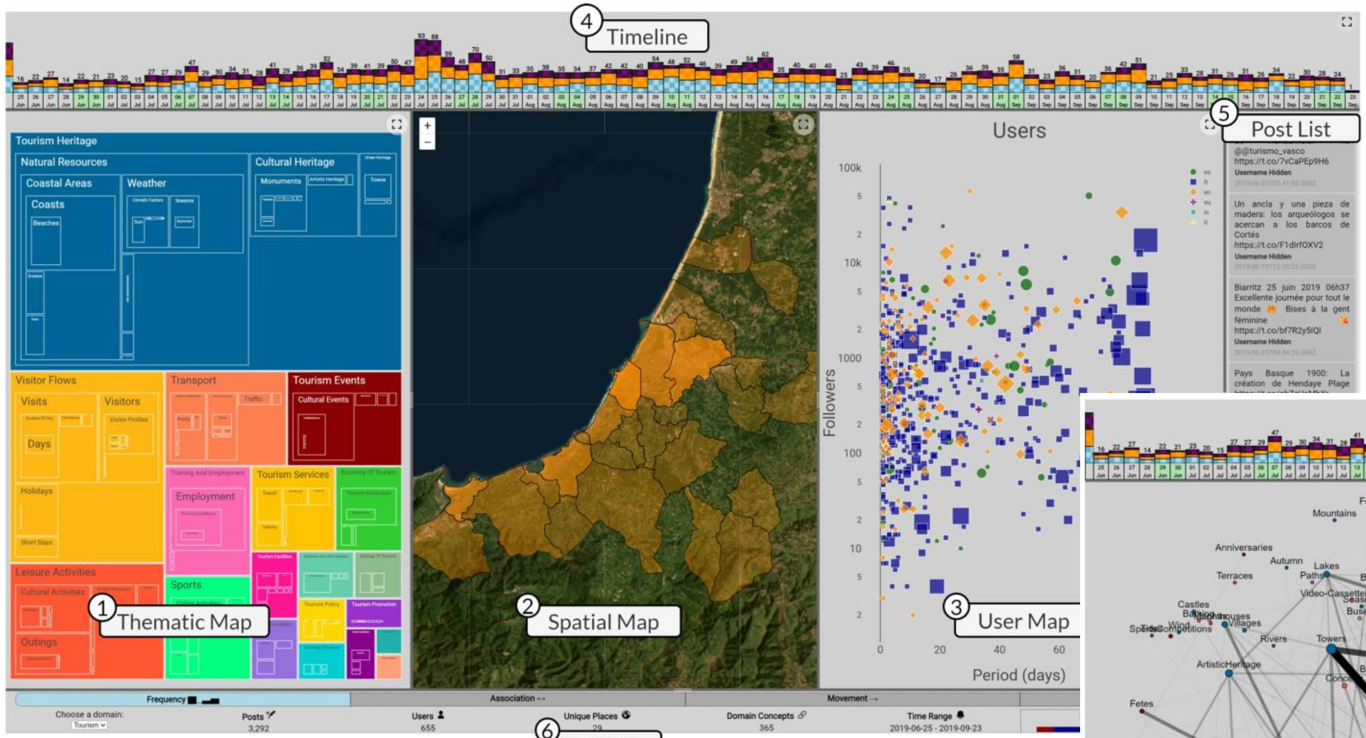
Conclusion

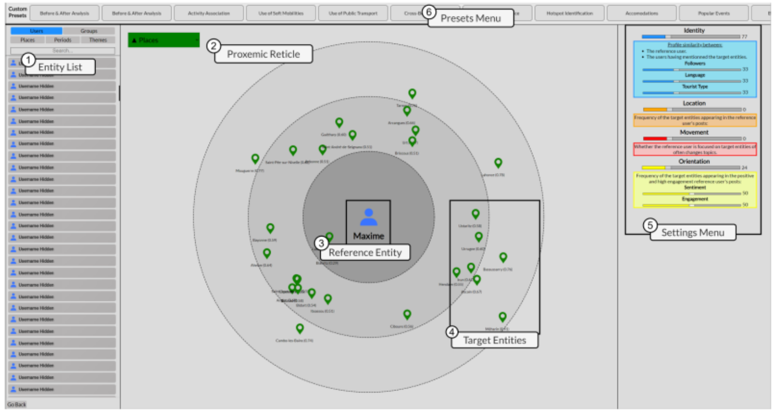
Perspectives

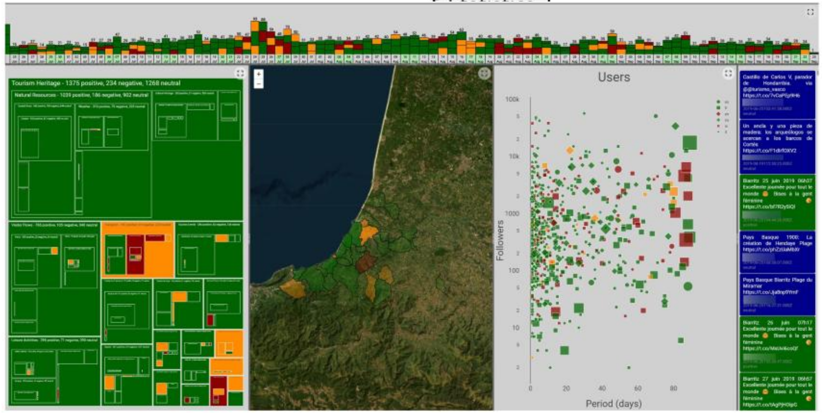
- Extension to **other application domains** to ensure the **generalizability** of our results
- Experiment on **larger dataset**
- Extension to **other languages** and **text types (e.g., newspaper)**
- Propose strategies for using LLMs for **the extraction of fine-grained thematic concepts**
  1. **High-level category processing** : group the 315 fine concepts into higher-level concepts for a first inference, then refine with the associated fine concepts
  2. **Batch processing** : divide the 315 fine concepts into small batches (e.g. 20) and run the inference in batches, then merge the results

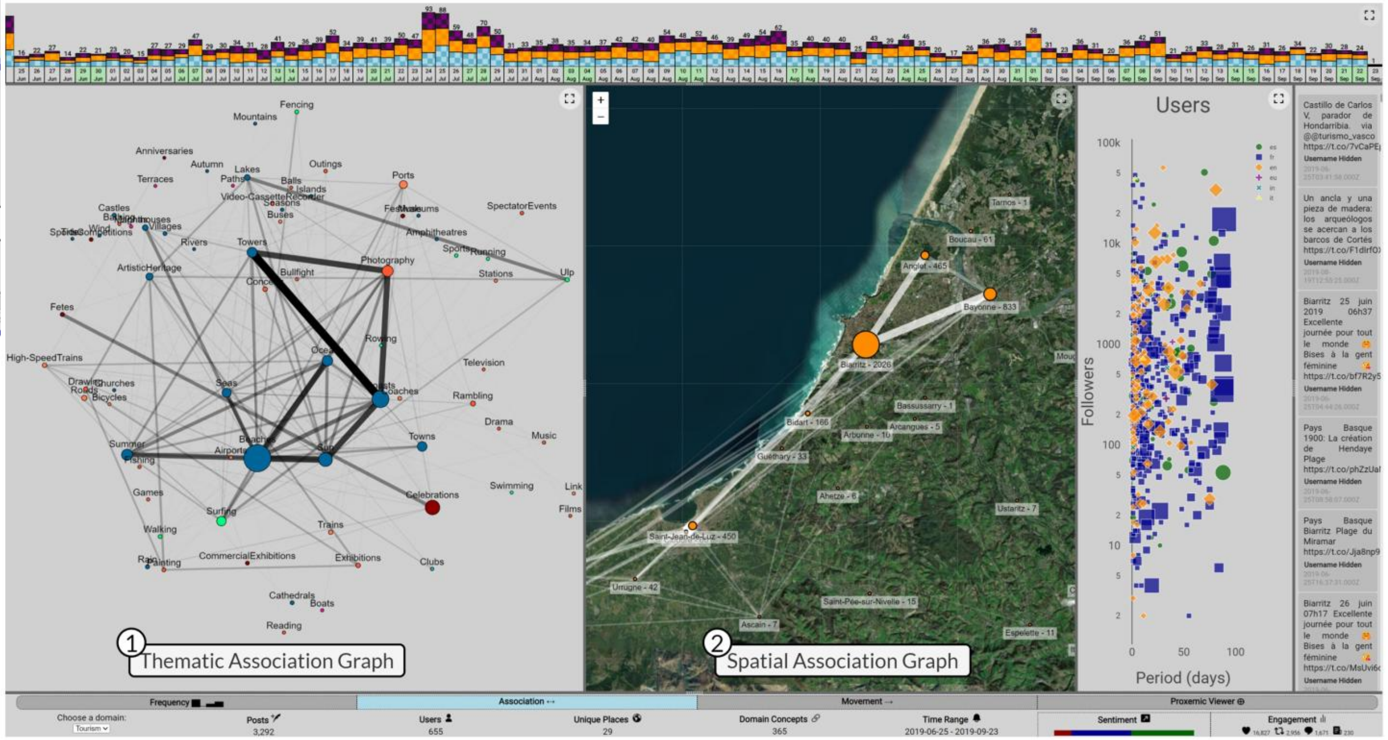
Conclusion

Example of application: The *TextBI* Dashboard









Thank you for your attention.

Any questions?