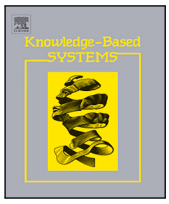
Knowledge-Based Systems 326 (2025) 114001

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Optimal strategies to perform multilingual analysis of social content for a novel dataset in the tourism domain

Maxime Masson [a,b,*], Rodrigo Agerri [b], Christian Sallaberry [a], Marie-Noelle Bessagnet [a], Annig Le Parc Lacayrelle [a], Philippe Roose [a]

[a] LIUPPA, E2S, University of Pau and Pays Adour (UPPA), France
[b] HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Spain

ARTICLE INFO

Keywords:
Tourism
Few-shot learning
Large language models
Masked language models
Multilinguality
Computational social science
Natural language processing

ABSTRACT

The rising influence of social media platforms in various domains, including tourism, has highlighted the growing need for efficient and automated Natural Language Processing (NLP) strategies to take advantage of this valuable resource. However, the transformation of multilingual, unstructured, and informal texts into structured knowledge still poses significant challenges, most notably the never-ending requirement for manually annotated data to train deep learning classifiers. In this work, we study different NLP techniques to establish the best ones to obtain competitive performances while keeping the need for training annotated data to a minimum. To do so, we built the first publicly available multilingual dataset (French, English, and Spanish) for the tourism domain, composed of tourism-related tweets. The dataset includes multilayered, manually revised annotations for Named Entity Recognition (NER) for Locations and Fine-grained Thematic Concepts Extraction mapped to the Thesaurus of Tourism and Leisure Activities of the World Tourism Organization, as well as for Sentiment Analysis at the tweet level. Extensive experimentation comparing various few-shot and fine-tuning techniques with modern language models demonstrate that modern few-shot techniques allow us to obtain competitive results for all three tasks with very little annotation data: 5 tweets per label (15 in total) for Sentiment Analysis, 30 tweets for Named Entity Recognition of Locations and 1K tweets annotated with fine-grained thematic concepts, a highly fine-grained sequence labeling task based on an inventory of 315 classes. We believe that our results, grounded in a novel dataset, pave the way for applying NLP to new domain-specific applications, reducing the need for manual annotations and circumventing the complexities of rule-based, ad-hoc solutions.

1. Introduction

Social media platforms have become essential channels for sharing opinions and experiences about tourist practices and itineraries. Twitter, in particular, has become a popular medium for people to spontaneously share their thoughts and recommendations, making it a valuable source of information and user-generated content (UGC) for the tourism industry [1] (such as tourism offices, destination marketing organizations - DMOs, etc.).

However, analyzing vast volumes of social media data can be challenging [2], especially when it comes to extracting structured knowledge from unstructured text. Consequently, tourism stakeholders often delegate the task of knowledge extraction to researchers, 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: Sentiment Analysis, NER for Locations, and Fine-grained Thematic Concept Extraction [3–5].

Recently, NLP techniques based on deep learning and language models have emerged. These techniques offer several advantages over traditional rule-based ones. Deep learning-based techniques can adapt to changing language patterns and structures [6], ensuring a more dynamic and up-to-date analysis. However, to achieve optimal results in domain-specific applications, language models must be fine-tuned. Consequently, researchers often face two recurrent challenges: (1) determining which NLP technique is most suitable for a given domain (e.g., which technique, which language model, etc.), and (2) discerning how many domain-specific examples are truly necessary to achieve competitive NLP results. As annotating datasets is both costly and

* Corresponding author at: LIUPPA, E2S, University of Pau and Pays Adour (UPPA), France.
E-mail addresses: maxime.masson@univ-pau.fr (M. Masson), rodrigo.agerri@ehu.eus (R. Agerri), christian.sallaberry@univ-pau.fr (C. Sallaberry), marie-noelle.bessagnet@univ-pau.fr (M.-N. Bessagnet), annig.lacayrelle@univ-pau.fr (A. Le Parc Lacayrelle), philippe.roose@univ-pau.fr (P. Roose).

https://doi.org/10.1016/j.knosys.2025.114001
Received 8 January 2024; Received in revised form 3 June 2025; Accepted 18 June 2025
Available online 1 July 2025

time-consuming, researchers strive to keep the annotation work to a minimum while maintaining high-quality results.

In this paper, we present a comparative study on the data requirements for achieving competitive performance in NLP tasks within the tourism domain. We hypothesize that among the existing language models and training approaches, some will be better suited for this specific domain, especially given the unique context of social media characterized by short informal texts, frequent grammatical errors, and the presence of emojis. We focus on the three common knowledge extraction tasks mentioned above: Sentiment Analysis (text classification), NER for Locations (sequence labeling), and Fine-grained Thematic Concept Extraction (sequence labeling). Specifically, we investigate which NLP techniques are best to keep manual data annotation to a minimum and to avoid cumbersome and costly rule-based ad-hoc approaches. To enable this study, we contribute a new manually annotated and multilingual (French, English, Spanish) dataset extracted from social media, which includes annotations for the three tasks mentioned above.

The main contributions are the following: (i) we present a novel dataset of tourism tweets. This dataset is multilingual (French, English, and Spanish) and has been manually annotated at the text level with sentiment labels and at the token level with locations and thematic concepts linked to a fine-grained tourism thesaurus (World Tourism Organization Thesaurus of Tourism and Leisure Activities [7]) which makes it, to the best of our knowledge, the first of its kind. The dataset is publicly available.[1]; (ii) we perform a comparative study on rule-based, fine-tuning, and few-shot learning techniques with the aim of establishing which of the techniques is the most efficient for each NLP task (Sentiment Analysis, NER for Locations and Fine-grained Thematic Concept Extraction) on social media data for the tourism domain; (iii) additionally, we experiment with various numbers of examples and dataset sampling techniques to determine how many annotated examples are really needed to achieve competitive results for each task, using different training techniques and language models.

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 1000 tweets annotated with thematic concepts. Our results also suggest that few-shot prompting for sequence labeling [8] 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 this type of task [9].

These results provide a promising solution to apply NLP to new domain-specific applications, keeping the manual annotation requirements low while avoiding complex rule-based ad-hoc solutions. The code and annotated data will be made publicly available to facilitate the reproducibility of the results and encourage research on this particular topic. In summary, we believe that the findings of our paper may be useful not only for researchers interested in the tourism domain but for any application that requires NLP analysis in a domain-specific scenario when no annotated data is available, while avoiding ad-hoc rule-based approaches.

The rest of the paper is structured as follows. In Section 2, we provide an overview of the NLP techniques based on deep learning used in tourism. Section 3 covers the construction and annotation of the dataset. Section 4 describes the experimental setup. In Section 5, we present a comparative analysis of fine-tuning and few-shot techniques for three common NLP tasks. Furthermore, the results and limitations of the experiment are discussed in Section 6. Finally, Section 7 provides some concluding remarks and offers some insight into the future application of the work presented here.

[1] https://huggingface.co/datasets/mx-phd/tourism.

Table 1
Examples of application of fine-tuning of language models in the tourism domain.

Ref.	Year	Objective	Label
[21]	2021	Spam detection (Hotel reviews)	Text
[22]	2022	Sentiment analysis (Tourism reviews)	Text
[23]	2021	Sentiment analysis (Tourism reviews)	Text
[24]	2021	Sentiment analysis (Transport)	Text
[28]	2018	Classification of Basque users	Text
[25]	2022	Location extraction (Touristic corpus)	Token
[26]	2020	Location extraction (Touristic corpus)	Token
[27]	2021	Travel themes identification	Token

2. Related work

In the rapidly evolving field of Natural Language Processing (NLP), one of the most significant advancements has been the advent of pre-trained language models based on the Transformer architecture [10]. These models are trained on vast corpora, capturing a broad spectrum of linguistic structures, nuances, and knowledge [6,11,12]. As a result, they offer a significant boost in performance and generalization for every NLP task. In this paper, we use all three 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 [10]. 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) [13] or XLM-RoBERTa [14].
- *Large Language Models* (LLMs) they 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 started to be used for discriminative tasks in few-shot settings [15–17]. Generative LLMs include the GPT (Generative Pre-Trained Transformer) series of models [18], Mistral [19], LLaMa 2 (Large Language Model Meta AI) series [20] or Google's FlanT5 [15] to name but a few.

2.1. Previous work on social media for the tourism domain

A popular approach in NLP is to fine-tune language models for domain-specific downstream tasks (refer to Table 1). 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 [21] and Sentiment Analysis in touristic reviews [22,23] or reviews about sustainable transport [24], leading to improved accuracy. In Named Entity Recognition (NER), fine-tuning of language models has been employed to extract location information from a tourism corpus [25,26]. Furthermore, language models have demonstrated promising results in thematic concept extraction, such as identifying travel-related themes and topics from tourism texts [27].

The main limitation of fine-tuning language models is that, in most cases, it requires substantial amount of manually annotated data to obtain competitive results [13]. These datasets are not always available, and sometimes they require being built from scratch, a costly and time-consuming process that researchers strive to avoid.

2.2. Addressing the lack of domain-specific annotated data

Zero-shot and *few-shot* learning techniques have emerged as effective approaches to mitigate the need for 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 us to obtain competitive results in classification tasks [29].

2.2.1. Few-shot with generative Large Language Models (LLMs)

In the case of generative LLMs (such as GPT [18], Mistral [19] or LLaMa 2 [20]), 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 [18]. However, results across domains are mixed. For example, studies have found that it can perform poorly in some domains, like the biomedical one [30].

2.2.2. Few-shot with Masked Language Models (MLMs)

With respect to text classification tasks, Pattern-Exploiting Training (PET) is a semi-supervised few-shot training approach that uses MLMs as a 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 [31]. 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 [9] 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. [9]. 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 [32] on the full training set of 3k examples [9].

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 [8,33]. For example, EntLM (Entity-oriented LM) [8] 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.

2.3. Cross-lingual transfer techniques

An alternative to few-shot learning 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 crosslingual 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 [34–36] and classification tasks [37,38] in various genres of text, including Sentiment Analysis in social media [39]. 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 of any annotated data in any language for touristic locations, fine-grained touristic concepts and sentiment for the tourism domain.

Table 2
Comparison of existing annotated datasets for various NLP tasks.

Dataset	Source	Annotations	Lang.	Type
ESTER [40]	Radio	Named entities	FR	Manual
AnCora [41]	Newspapers	Named entities	ES, CA	Semi
BTC [42]	Twitter	Named entities	EN	Manual
Sent140 [43]	Twitter	Sentiment	EN	Auto
STS-Gold [44]	Twitter	Sentiment	EN	Manual
GSC [24]	Reviews	Sentiment	EN	Auto
SemEval [45]	Twitter	Sentiment	EN	Manual
MultiWOZ [49]	Humans	Dialogue states	EN	Manual
SNLI [50]	Humans	Inference pairs	EN	Manual
Heldugazte [28]	Twitter	Formal/Informal	EU	Auto

2.4. Existing annotated resources

While publicly available annotated data for the tourism domain is non-existent, there are several annotated corpora from other domains that could be used for experimentation, as highlighted in Table 2. Here, we compare a selection of existing annotated datasets along different criteria: (1) the source of the data collection, (2) the types of annotations available, (3) the languages covered by the dataset, and (4) the methodology used for generating annotations (manual by humans, semi-automatic or automatic).

For example, the ESTER corpus [40] is a comprehensive collection of French radio transcripts and AnCora [41], is a multilevel annotated corpus (mostly from newspaper) for Catalan and Spanish. Both of these resources are annotated for named entities (such as persons, locations, organizations).

In terms of social media-specific resources, the Broad Twitter Corpus (BTC) [42] includes coarse-grained NER annotations, while Sentiment140 [43], STS-Gold [44], and many other datasets developed as part of shared evaluation tasks at SemEval [3,45–48], are used for sentiment analysis. Other corpora include the MultiWOZ dialogue dataset [49], the Stanford NLI dataset [50] for text inference, and the Heldugazte corpus [28], which assists in categorizing tweets as formal or informal.

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. Dataset building and annotation

In this section, we describe the creation of a novel multilingual dataset[2] consisting of tourism-related tweets annotated for three important NLP tasks for tourist applications: (1) Sentiment Analysis, (2) Named Entity Recognition for Locations, and (3) Fine-grained Thematic Concept Extraction (based on the World Tourism Organization Thesaurus on Tourism and Leisure Activities [7]).

3.1. Collection

The dataset was collected from Twitter using the Academic API[3] and the collection process was carried out by applying a novel methodology that we have been designed for dataset building. This methodology is both generic, iterative and incremental [51]. Several iterations were carried out, each of them with successive filtering corresponding to the dimensions of the target dataset:

[2] https://huggingface.co/datasets/mx-phd/tourism.

[3] https://developer.twitter.com/en/use-cases/do-research/academic-research (discontinued in April 2023).

Table 3
Breakdown of the collected dataset by language – tweets (Users).

	All	French	English	Spanish
Train	1662 (503)	1297 (391)	283 (129)	82 (32)
Dev	619 (300)	450 (213)	99 (66)	70 (31)
Test	680 (431)	401 (273)	102 (100)	177 (93)

- *Spatial*: French Basque Coast area (defined by spatial coordinates for geotagged tweets or a list of toponyms).
- *Temporal*: Summer of 2019 (21 Jun – 21 Sept). Based on tweets' timestamps.
- *Thematic*: Tourism domain (defined by the World Tourism Organization Thesaurus on Tourism and Leisure).

To prevent excessive noise, each iteration was followed by human feedback to adjust and balance the filters. For example, a large number of tweets related to the G7 summit, which took place in the region that year, were collected because they contained tourism-related keywords; these were subsequently blacklisted using hashtag and keyword exclusion rules. Similarly, we excluded professional and institutional users because we are primarily interested in understanding the behaviors and feelings of individual tourists, not analyzing promotional or institutional content.

The data collection strictly adhered to Twitter's Academic Research API policy, which, at the time of collection, authorized the retrieval of up to 10 million tweets per month for non-commercial research. In accordance with GDPR and standard ethical practices, only tweet texts and user IDs were retained; all personal metadata, including usernames and profile information, were removed to ensure user anonymization and privacy protection.

The final dataset is made up of 27,379 tweets, of which 2961 tweets from 624 users were selected to be annotated and used for experimentation. The tweets in this subset (2961) were manually checked to confirm that they were about tourism and from tourist visitors (unlike the tweets from tourism professionals, or from news outlets speaking about tourism, etc.). The dataset is multilingual and includes an unbalanced variety of tweets in English, French, and Spanish, which reflects the reality of the use of social media on the French Basque Coast. Table 3 provides the language distribution of the dataset and the splits created for experimentation (60% train, 20% dev, 20% test). Splits were generated maintaining a balance between the number of users and languages in each set.

3.2. Sentiment annotation

The annotation process of those 2961 tweets was carried out in a semi-automatic manner, following the procedure depicted in Fig. 1.

Firstly, to assist human annotators, the 1299 tweets in the development and test splits underwent a process of automatic annotation using the 5 language models listed in Table 4. Subsequently, they were manually reviewed. Each tweet was assigned to two annotators to evaluate agreement (Cohen's kappa coefficient) 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 the human annotations. Table 4 shows that XLM-T Sentiment, fine-tuned with multilingual Sentiment Analysis data from various domains different from tourism [52], outperformed, on average for the three languages, any other approaches. Following this, we automatically annotated using XLM-T Sentiment the training split, which was then manually revised for its use in the experimentation.

Table 4
Accuracy of available sentiment language models on manually annotated test data [53–56].

Sentiment models:	Barbieri et al. (2020) [53]	Pérez and Carlos (2021) [54]	Seethal (2022) [55]	Hartmann et al. (2023) [56]	Barbieri et al. (2022) [52]
French					
Global	0.56	0.45	0.43	0.47	**0.82**
Positive	0.34	0.14	0.11	0.95	0.82
Negative	0.06	0.11	0.00	0.28	1.00
Neutral	0.97	0.97	1.00	0.00	0.74
Spanish					
Global	0.71	0.64	0.61	0.34	**0.83**
Positive	0.31	0.09	0.03	1.00	0.84
Negative	0.00	0.00	0.00	0.29	0.43
Neutral	1.00	1.00	0.98	0.00	0.87
English					
Global	**0.81**	**0.81**	0.71	0.66	0.80
Positive	0.75	0.75	0.59	1.00	0.72
Negative	0.75	0.50	0.75	0.50	0.75
Neutral	0.94	0.97	0.94	0.00	0.97

3.3. Locations and thematic concepts

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 thematic concepts relevant to the tourism domain. These two tasks are addressed as sequence labeling problems.

Before experimenting with supervised techniques, we implemented a basic rule-based *word-matching* approach as a baseline for locations and thematic concepts. Locations were matched using 625 local toponyms extracted from Open Street Map (cities, POIs, landmarks, etc.) while thematic concepts were matched using their label and synonyms in the WTO thesaurus of tourism (which contains 1494 touristic concepts).

Tweet 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, all train, dev, and test sets. For thematic concepts the *word-matching* algorithm detected 315 concept classes for the full dataset (out of the 1494 concepts included in the WTO ontology), making it a highly fine-grained sequence labeling task. Due to this fact, we just revised the 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)). 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 *word-matching* baseline on the test set are reported in Table 5. These results will serve as a baseline to compare with the different supervised techniques in the experimental section.

From the results in Table 5, it can be observed that, for locations, results are not that good, 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

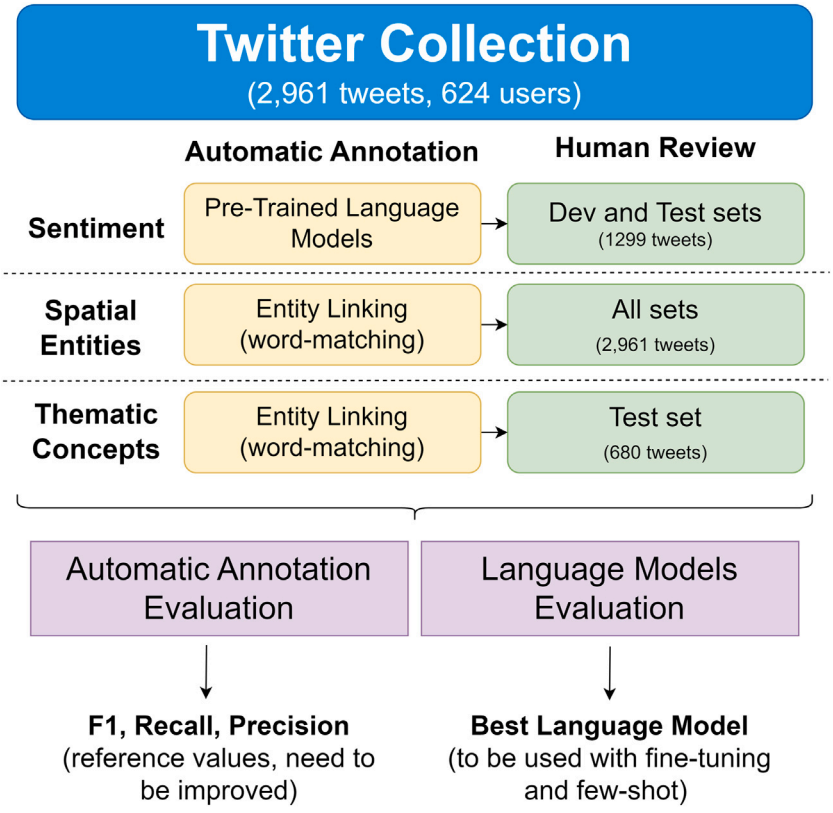


Fig. 1. Dataset building and annotation process.

Table 5
Performance of the *word-matching* algorithm on both sequence labeling tasks.

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 6
Dataset annotations (following human review).

Set	Tweets	Locations	Concepts	Sentiments
Train	1662	4030	3841	787 (+) 191 (–) 684 (=)
Dev	619	1419	1337	271 (+) 82 (–) 266 (=)
Test	680	1679	1844	299 (+) 93 (–) 288 (=)
Global	2961	7128	7022	1357 (+) 366 (–) 1238 (=)

this constitutes a rather strong baseline, the recall remains comparatively low, which means that many thematic concepts remain undetected by the system. Furthermore, developing such as *word-matching* algorithm is a rather complex and time-consuming exercise which we would ideally like to avoid for any domain-specific new application.

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 techniques, it is of particular interest to investigate techniques based on few-shot learning, 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 6 provides a detailed description of the dataset, including the total number of tweets (2961) and the number of annotations per task. Note that for the sentiment column in Table 6, the colors and symbols used represent each type of polarity (– for negative annotations, + for positive and = for neutral).

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

4. Experimental setup

Fig. 2 provides an overview of our experimental setup with the language models, sampling techniques, and learning techniques used for each task. The experimentation is focused on the three tasks presented above: Sentiment Analysis (see Section 4.1), Named Entity Recognition (NER) for Locations and Fine-grained Thematic Concept Extraction (see Section 4.2). These experiments leverage the tourist dataset described in the previous section (2961 multilingual tweets including 1662 for training) and summarized in Table 6. Table 7 provides an overview of the hyperparameters used.

To study machine learning approaches, the dataset is sampled using two different techniques (see Fig. 2, *Dataset Sampling Techniques*):

- *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 [31] while for locations and fine-grained thematic concepts, we apply the EntLM *k*-shot technique with default parameters [8].

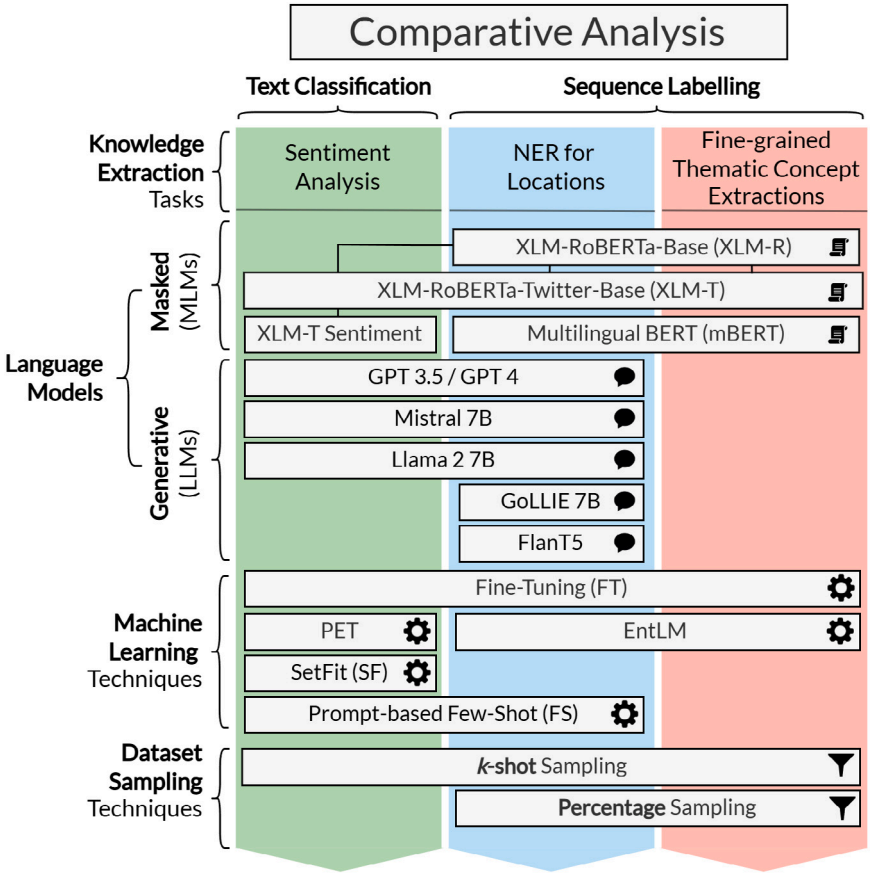


Fig. 2. Experimental setup of the comparative study.

- *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 which are neither location nor thematic entities), which the k-shot sampling technique does not contemplate.

4.1. Text classification – Sentiment analysis

Firstly, based on the results reported in Table 4, we used MLMs (see Fig. 2, *Language Models, Masked*). We chose the XLM-T for experiments on Sentiment Analysis. XLM-T is based on XLM-RoBERTa [14], further pre-trained on a corpus of 198 million tweets for 15 languages [52]. 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 [52] (XLM-T).
- XLM-T was previously fine-tuned specifically for Sentiment Analysis [52] (XLM-T Sentiment). This sentiment variant has already been 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, which do not necessarily include tourism.

We also experiment with the following generative LLMs (see Fig. 2, *Language Models, Generative*):

- GPT 3.5 [18] (gpt-3.5-turbo-0125),[4] 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.[5] Additionally, we also use GPT 4 (gpt-4-0125-preview) but only in zero-shot settings due to API cost limitations. For both models, we use the default temperature of 1 and a presence penalty of 0.
- Mistral 7B [19]: 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 [57].
- LLaMa 2 7B [20]: 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 [58].

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 Section 2.2. 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 [57,58].

With respect to the two MLMs (e.g., XLM-T and XLM-T Sentiment), we used them as a backbone for three different training techniques (see Fig. 2, *Machine Learning Techniques*) along various dataset sizes using the sampling techniques described previously.

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

- *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) [31] is used with default hyperparameters.
- *SetFit* (SF) [9]: a prompt-free framework for few-shot fine-tuning of sentence transformers. We use the training parameters recommended in the SetFit repository,[6] namely 4 epochs of training with a batch size of 16.

By comparing these deep learning 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.

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.
- *EntLM* [8] with a multilingual BERT (mBERT)[13] 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 [14] (XLM-R) and the previously mentioned mBERT. Grid search for hyperparameter tuning found as 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.[16]. 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 [17]. 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 learning 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* (rules-based) algorithm, as presented in Table 5. The primary objective of this comparison is to ascertain the minimum amount of annotated data required to justify transitioning from costly and manual *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, which involves labeling 315 different concept classes, and for which developing *word-matching* algorithms or manually annotating data are highly inefficient and expensive approaches.

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

Table 7
Overview of the hyperparameters used for each approach.

Technique	Batch	Learning rate	Decay
Text classification			
PET	Default (Schick et al. 2020) [31]		
SetFit (SF)	16	4 epochs	
Fine-tuning (XLM-T Sent)	32	1e−5	0.1
Fine-tuning (Mistral/LLaMa 2)	1	2e−4	0.001
Sequence labeling			
EntLM (mBERT)	4	1e−4	0
GoLLIE	Default (Sainz et al. 2024) [17]		
Fine-tuning (XLM-R/mBERT)	8	5e−5	0.1
Fine-tuning (FlanT5/LLaMa 2)	As per García-Ferrero et al. (2024) [16]		
Both			
Fine-tuning (XML-T)	8	2e−5	0.01
Few-shot (GPT-3.5/GPT-4)	Temp: 1, Presence Penalty: 0		
Few-shot (Mistral/LLaMa 2)	Default [19,20]		

To handle noise in tweets (e.g., spelling errors, slang), we rely on the inherent robustness of pretrained language models, which have been trained on large-scale, noisy corpora including social media text. These models generalize well to informal language, spelling variations, and slang, eliminating the need for additional noise-specific preprocessing. Their learned representations implicitly normalize such variability. For related work demonstrating language models' robustness to noise, see [18,59].

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 [60].

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 [61]. All reported results are the average of three randomly initialized runs.

5. Results

We apply the experimental setup presented in the previous section to our multilingual dataset of tweets from the tourism domain.

5.1. Sentiment analysis

The results of the Sentiment Analysis on the five techniques are reported in Table 8. 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 8, *Fine-Tune of MLMs*) outperforms any other method using only 5 examples for training (Table 8, 0.919). In contrast to previous work [31], 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 tourist data, clearly outperforming few-shot prompting techniques with

Table 8
Sentiment analysis with *k*-shot sampling - Results on text classification techniques (results in **bold** are referenced in the text).

	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 Sent		**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 Sent.		0.831	0.878	0.876	0.893	0.882	0.899	0.858	0.821

MLMs such as PET [31] or LLMs. In fact, the fine-tuned XLM-T Sentiment model reaches optimal results with as few as 10 examples (Table 8, 0.939).

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).

Concerning 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 8), but with more data, results from PET and fine-tuning are eventually better. Still, this highlights the effectiveness of SetFit, which can achieve high accuracy with only 40 examples and without requiring as many computing resources as for few-shot learning with LLMs.

The performance of large language models (LLMs) in the zero-shot setting (Table 8, *Prompt-based Few-Shot*) shows that GPT 3.5 and Mistral 7B achieve competitive accuracy, with results ranging from 0.716 for Mistral 7B to 0.785 for GPT 3.5. In our experiments, GPT 3.5 outperformed GPT 4, achieving an accuracy of 0.785 compared to 0.757 for GPT 4. Due to the high cost associated with the GPT 4 API, higher shot configurations were not tested, and as a result, GPT 4's performance in those settings is not reported in Table 8.

The observed difference between GPT 3.5 and 4 may be attributed to the different underlying pre-training and post-training optimizations. Moreover, the performance of Mistral 7B, though slightly lower than GPT 3.5, is still competitive, which suggests that open-source models can achieve near-state-of-the-art results while being more cost-efficient.

Fig. 3 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 Fig. 3, 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 Fig. 3 and Table 8 suggest 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 8, *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 Fig. 3, 40 examples).

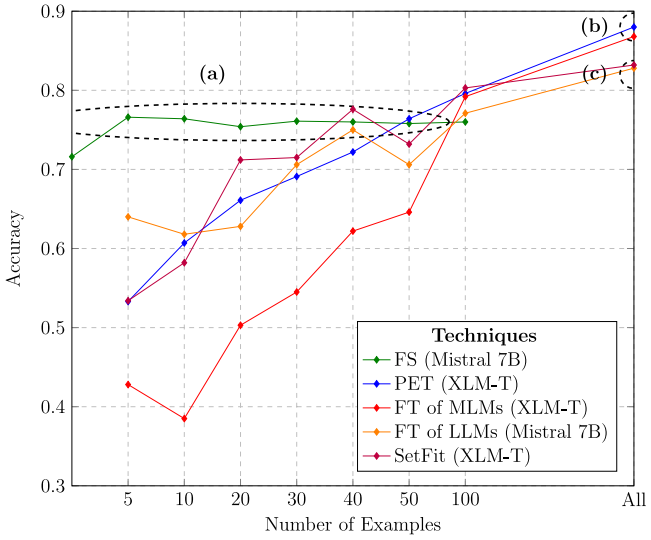


Fig. 3. Sentiment analysis - *k*-shot sampling.

3. In use cases where very little 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 approach. This approach is able to produce an accuracy of roughly 0.750 consistently from 5 to 100 examples (refer to Fig. 3, (a)). 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 sizable amount of annotated data is available, Pattern-Exploiting Training and Fine-Tuning of MLMs (refer to Fig. 3, (b)) produce slightly better results in the tourism domain than SetFit and Fine-Tuning of LLMs (refer to Fig. 3). 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 6. In the next section, we present the results obtained for the sequence labeling tasks, namely NER for Locations and Fine-grained Thematic Concept Extraction.

5.2. Named Entity Recognition (NER) for locations

Table 9 reports the results of NER for Locations. When the full training dataset is employed, all techniques yield comparable F1-scores (refer to Table 9, *All Examples*).

Table 9
Named Entity Recognition (NER) for locations with *k*-shot sampling - Results on sequence labeling techniques (results in **bold** are referenced in the text).

	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**

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's few-shot scores are slightly higher, Mistral is an open-weights 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 9). 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.

Fig. 4 provides a chart view of the *k*-shot sampling results. For each technique in Table 9, we report in the chart the results obtained with the most efficient open-source language model. Summarizing, several key takeaways can be drawn from Fig. 4:

1. With few examples, few-shot learning with LLMs, such as GPT 3.5, produces the by far the best results. It matches the *word-matching* approach (reference F1-score) in zero-shot settings and surpasses 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 recurrent 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 learning is beginning to shift this paradigm.
3. EntLM with MLMs yields more modest results but remains viable in low-shot settings, unlike fine-tuning techniques.

Fig. 5 shows selected techniques used with percentage sampling (LLMs did not fit due to context 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

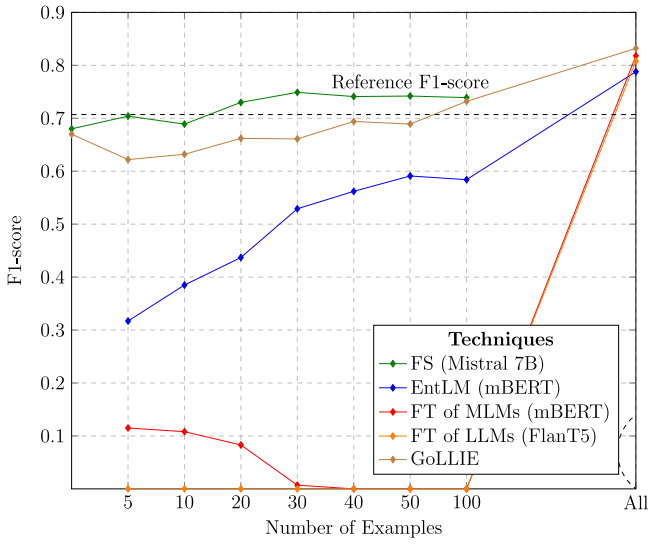


Fig. 4. Named Entity Recognition (NER) for locations – *k*-shot sampling.

the way this task is set up 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 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. An interesting finding of our experiments (see Fig. 5) is that both fine-tuned mBERT and EntLM are superior to the *word-matching* algorithm when using only approximately 1000 tweets for training.

As depicted in Fig. 5, approximately 20% of the dataset (equivalent to about 330 tweets in our study) is necessary to achieve competitive results using this method. 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.

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 Fig. 5, 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 9, 0.584 with 100 examples).

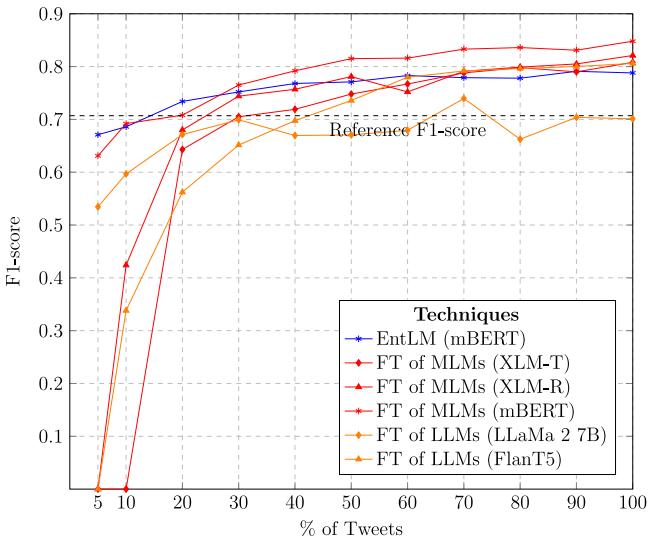


Fig. 5. Named Entity Recognition (NER) for locations – Percentage sampling.

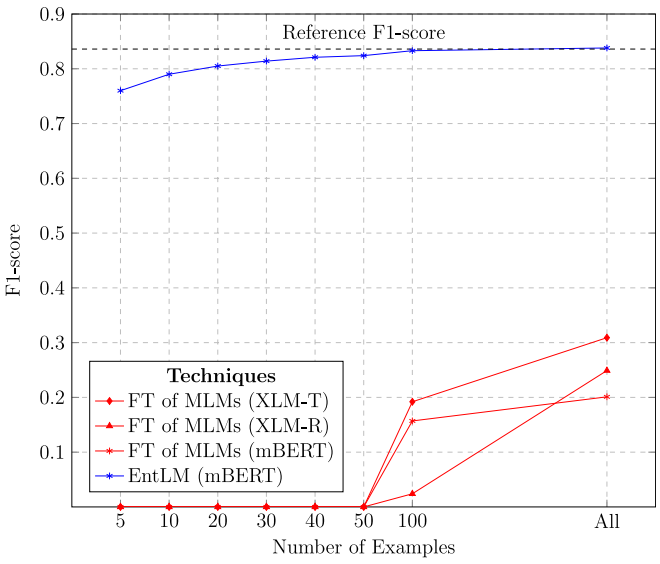


Fig. 6. Fine-grained thematic concept extraction – k-shot sampling.

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 learning techniques with LLMs excel (refer to Table 9, *Prompt-based Few-Shot*). Particularly in zero-shot settings, the best results are achieved with GPT 4, which produces a remarkable F1-score of 0.829, equaling or surpassing all other strategies, even those involving extensive 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.

5.3. Fine-grained thematic concept extraction

Perhaps it is in the evaluation of Fine-grained Thematic Concept Extraction, shown in Figs. 6 and 7, where few-shot learning 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. 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 effort 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.

Let us now move on to case studies illustrating the difficulty of the tasks.

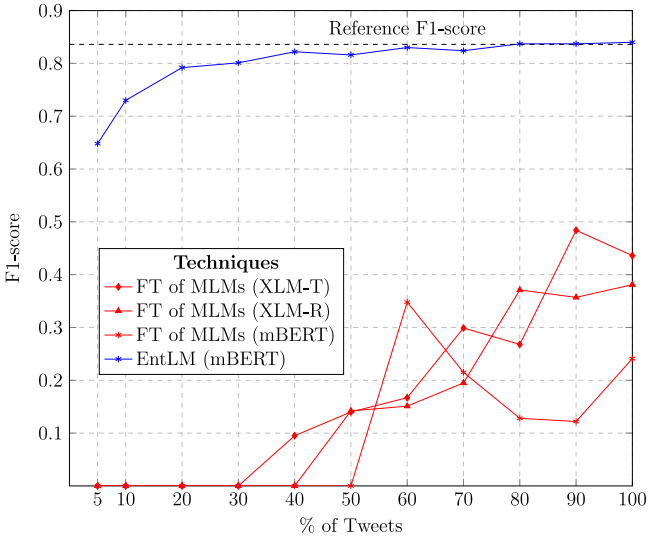


Fig. 7. Fine-grained thematic concept extraction – Percentage sampling.

5.4. Case study: Visualization of the results for the tourism domain

Fig. 8 shows a visualization of the results generated using a dashboard that was created as part of our project to visualize NLP annotations, see [62] for more details. Here, we have loaded the annotations generated by the best techniques for each task.

Firstly, on the left, we have a thematic map (represented as a multilevel treemap). Here, each square represents a thematic concept, and the size of each square represents the proportion of the concept in the dataset. As we can see, *TourismHeritage*-related concepts are in the majority (e.g., *NaturalResources*, *CulturalHeritage*). Additionally, we have superimposed a sentiment overlay on the thematic map to visualize the aggregated sentiment for each thematic concept. Here, green signifies that the sentiment tends to be positive, orange mixed, and red negative. We can observe that at the scale of our dataset, most touristic concepts tend to be associated with a positive sentiment, except some branches like *Transportation*, which are more mixed and even plainly negative for some concepts.

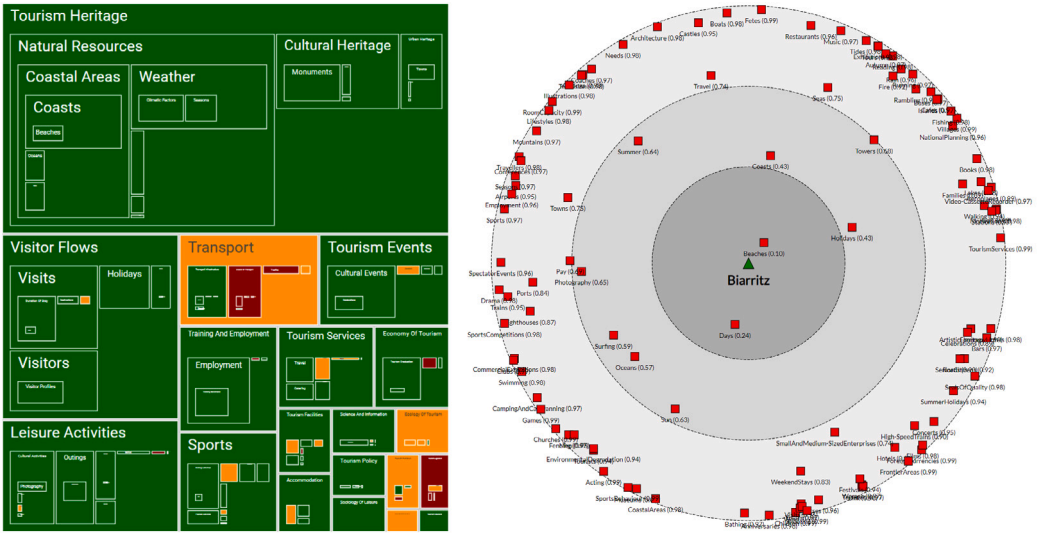


Fig. 8. Example of NLP results usage – Tourism thematic map with sentiment overlay (left) and proxemic view (right).

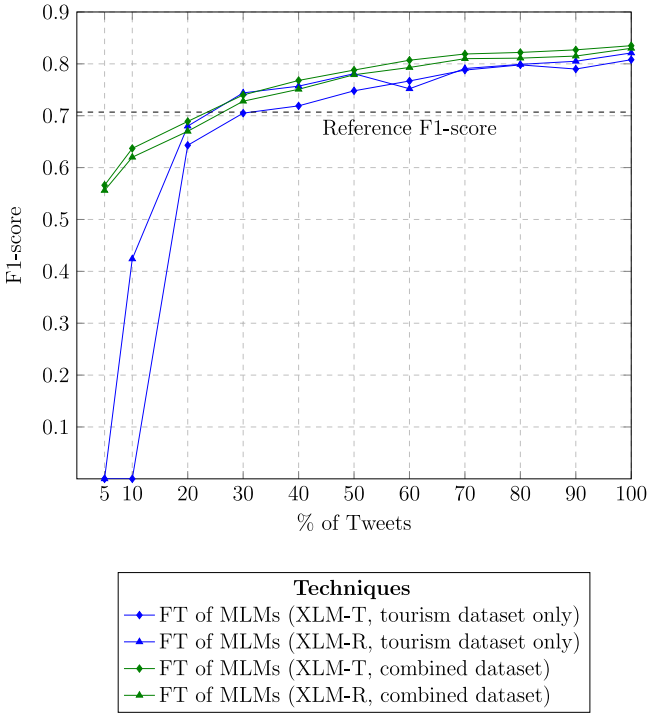


Fig. 9. Comparison of Named Entity Recognition (NER) model performance: Our dataset (blue) compared to the combined dataset (green) – fine-tuning.

On the right is a proxemic crosshair [63]. Here, a location named entity (the city of *Biarritz*) is represented at the center, and the most co-occurring thematic concepts are scattered around it. This allows us to visualize for a given location, which touristic concepts it is mostly associated with. In this case, *Biarritz* is mostly associated with the concepts of *Beaches*, *Holidays*, *Coasts*, *Days*, and *Ocean*.

This visualization illustrates the potential of this kind of NLP annotations for stakeholders in the tourism domain (e.g., tourism offices, local municipalities, etc.). 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.

6. Discussion and error analysis

In our comparative analysis of various techniques for three knowledge extraction tasks in the tourism domain, namely Sentiment Analysis, NER for Locations, and Fine-grained Thematic Concept Extraction, we have gained interesting insights into the data requirements and performance of these techniques.

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 learning techniques with LLMs (in particular the GPT series of models and Mistral 7B) has been demonstrated to be the best technique to avoid costly manual annotation work.

As for misclassification cases, they often arise from the model's reliance on surface-level lexical features without adequately accounting for context or tone. For example, the tweet "*Feu d'artifice du 14 juillet à Hendaye*",[7] referring to a national celebration, is incorrectly labeled as negative by XLM-T Sentiment despite its neutral or festive nature. Similarly, "*Balade à Hendaye #plage #mer #architecture*"[8] describes a leisure activity in a neutral setting, but is misclassified, likely due to the absence of explicit sentiment words and overemphasis on structural patterns.

6.2. Named Entity Recognition (NER) for locations

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

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 that associating many label words with the same class does not benefit EntLM.

Interestingly, prompt-based few-shot learning with LLMs and GoLLIE generally performs exceptionally well in contexts with limited

[7] English version: "July 14th fireworks in Hendaye".
[8] English version: "Walk in Hendaye #beach #sea #architecture".

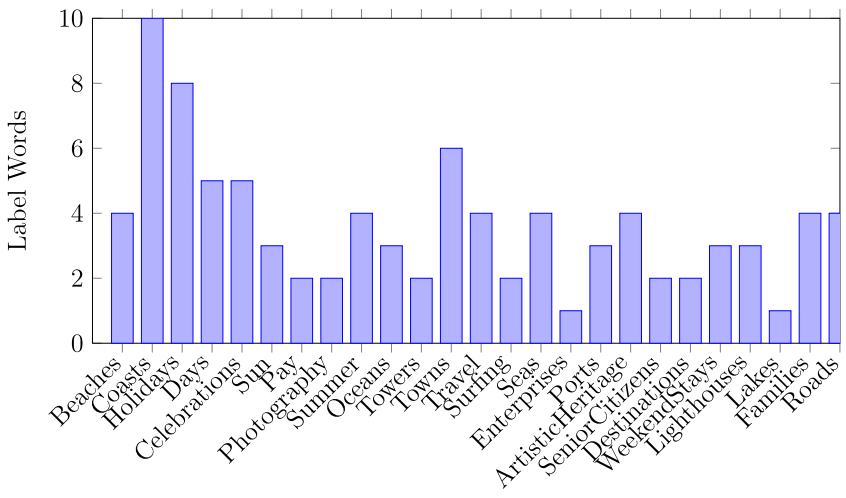


Fig. 10. Number of label words for the most frequent thematic concepts found in the tweets.

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.

Regarding the misclassification cases in NER for Locations, they frequently result from ambiguity in sequence boundaries, improper capitalization, or domain-specific phrasing that deviates from standard named entity patterns. For instance, a phrase like "congrès #sniteat-unsa à Hendaye"[9] may lead the models to misidentify "#sniteat-unsa" as a location rather than an organization (SNITEAT-UNSA is a French trade union) due to the hashtag and its syntactic position. Similarly, "Plage d'Hendaye"[10] might be incorrectly labeled as a facility instead of a geographic location. These cases illustrate challenges in handling informal structures, social media conventions, and multilingual inputs.

We also explored the idea of improving the fine-tuning process of our dataset by combining it with other existing corpora from other domains (see Fig. 9), such as AnCora [41] (Spanish) and ESTER [40] (French), both already annotated with location entities. However, this experiment did not lead to any significant improvements in the F1-score as shown in Fig. 9. 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. 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 consists of newspaper texts. Consequently, they lack the contextual information pertinent to the tourism domain.

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, which involves a large inventory of classes (315 concepts instantiated out of the 1494 from the WTO tourism thesaurus), each having very few instance label words that are highly representative of the classes to which they refer. Fig. 10 shows the low count of unique label words for the thematic concepts most often 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 of this paper is that the EntLM few-shot learning 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 with each of the classes, namely, the thematic concepts. This result is particularly useful for practical applications where manually annotated data is usually very scarce.

We also performed a manual inspection of the misclassification cases in thematic concept detection, errors are largely due to the high number of distinct themes combined with sparse training instances per theme. This imbalance hinders the models' ability to generalize, leading to poor recognition of underrepresented concepts. Additionally, the models often fail to detect concepts when they are misspelled, abbreviated, or expressed in non-standard forms: e.g., "adminstratif" instead of "administratif" or hashtags like "#contentieuxadministratif"[11] being overlooked due to their concatenated structure. These factors reduce robustness in real-world, noisy text environments.

Our experiments provide interesting insights but have some limitations. Firstly, we intentionally focused on three common NLP 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 2961 tweets. This limitation in dataset size was purposefully chosen 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 tourist demographics of the French Basque Coast region.

7. Concluding remarks and future work

In this paper, we present a comparative study of several Natural Language Processing (NLP) strategies for Sentiment Analysis, NER for Locations, and Fine-grained Thematic Concept Extraction for social media data in the tourism domain.

[9] English version: "conference #sniteat-unsa in Hendaye".
[10] English version: "Hendaye Beach".
[11] English version: "#administrativelitigation".

The main objective of this work is to establish the best approaches to obtain competitive results while keeping to a minimum any costly manual annotation or the development of complex and cumbersome rule-based approaches, especially for complex tasks such as Fine-grained Thematic Concept Extraction. In order to be able to do the experimentation, we provide a novel tourist-specific multilingual (French, English, and Spanish) dataset annotated for the tasks that are the subject of our study. This dataset, one of its kind, will be made publicly available in the future to facilitate further research on this particular topic and the reproducibility of the results.

Results show that current few-shot learning techniques allow us to obtain competitive results for all three tasks with a very small amount of annotation: 5 tweets per label (15 in total) for Sentiment Analysis, 30 examples of Location for NER, and 1000 tweets annotated with thematic concepts. We believe that these findings are helpful not only for the development of our own application but also for other domain-specific applications, which may require NLP analysis as a model enrichment. However, further research on analogous NLP tasks in different domains would be needed to validate the generalizability of our results in various domains, other than tourism. We believe these findings offer value not only for NLP researchers focusing on tourism but also for applications needing domain-specific NLP analysis, especially when there is a lack of annotated data or when one wishes to exclude ad hoc rule-based approaches.

Our project's next step involves presenting the NLP results to stakeholders in the tourism industry. To achieve this, we have initiated the development of multidimensional, dynamic dashboards based on the output generated by the NLP processing modules discussed in this paper. These dashboards are designed to highlight the frequency, relationships, and trajectories of extracted entities, categorized spatially (cities, POIs), temporally (date, time periods), and thematically (touristic concepts), within a specified social media corpus, all correlated with contextual data such as sentiment and engagement levels.

CRediT authorship contribution statement

Maxime Masson: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rodrigo Agerri:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Christian Sallaberry:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Marie-Noelle Bessagnet:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Annig Le Parc Lacayrelle:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Philippe Roose:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the urban community of Pau Béarn Pyrénés and E2S UPPA.

Data availability

Data will be made available on request.

References

[1] B. Zeng, R. Gerritsen, What do we know about social media in tourism? A review, Tour. Manag. Perspect. 10 (2014) 27–36.
[2] D. Maynard, K. Bontcheva, D. Rout, Challenges in developing opinion mining tools for social media, in: Workshop Programme, p. 15.
[3] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, V. Stoyanov, SemEval-2015 task 10: Sentiment analysis in Twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEVal 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 451–463, http://dx.doi.org/10.18653/v1/S15-2078, URL https://aclanthology.org/S15-2078.
[4] X. Liu, H. Chen, W. Xia, Overview of named entity recognition, J. Contemp. Educ. Res. 6 (5) (2022) 65–68.
[5] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K.J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, et al., Clinical concept extraction: a methodology review, J. Biomed. Inform. 109 (2020) 103526.
[6] B. Min, H. Ross, E. Sulem, A.P.B. Veyseh, T.H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Comput. Surv. 56 (2) (2023) 1–40.
[7] World Tourism Organization, Thesaurus on Tourism and Leisure Activities 2001, World Tourism Organization, Madrid, Spain, 2002.
[8] R. Ma, X. Zhou, T. Gui, Y. Tan, L. Li, Q. Zhang, X. Huang, 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, Association for Computational Linguistics, 2022, pp. 5721–5732, http://dx.doi.org/10.18653/v1/2022.naacl-main.420.
[9] L. Tunstall, N. Reimers, U.E.S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, 2022, http://dx.doi.org/10.48550/ARXIV.2209.11055, URL https://arxiv.org/abs/2209.11055.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[11] C.D. Manning, K. Clark, J. Hewitt, U. Khandelwal, O. Levy, Emergent linguistic structure in artificial neural networks trained by self-supervision, Proc. Natl. Acad. Sci. 117 (2020) 30046–30054, http://dx.doi.org/10.1073/pnas.1907367117.
[12] O. Toporkov, R. Agerri, On the role of morphological information for contextual lemmatization, Comput. Linguist. (2024) 1–35.
[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, CoRR, abs/1810.04805 arXiv:1810.04805 URL http://arxiv.org/abs/1810.04805.
[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
[15] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, J. Mach. Learn. Res. 25 (70) (2024) 1–53.
[16] I. García-Ferrero, R. Agerri, A.A. Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, et al., Medical mT5: An open-source multilingual Text-to-Text LLM for the medical domain, 2024, arXiv preprint arXiv:2404.07613.
[17] O. Sainz, I. García-Ferrero, R. Agerri, O.L. de Lacalle, G. Rigau, E. Agirre, GoLLIE: Annotation guidelines improve zero-shot information-extraction, in: The Twelfth International Conference on Learning Representations, 2024, URL https://openreview.net/forum?id=Y3wpuxd7u9.
[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[19] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7B, 2023, arXiv preprint arXiv:2310.06825.
[20] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, 2023, arXiv preprint arXiv:2307.09288.
[21] M. Crawford, T.M. Khoshgoftaar, 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, IEEE, 2021, pp. 248–254.
[22] M.P. Enríquez, J.A. Mencía, I. Segura-Bedmar, Transformers approach for sentiment analysis: Classification of Mexican tourists reviews from TripAdvisor, 2022.
[23] J. Vásquez, H. Gómez-Adorno, G. Bel-Enguix, Bert-based approach for sentiment analysis of spanish reviews from TripAdvisor, in: IberLEF@ SEPLN, 2021, pp. 165–170.

[24] A. Serna, A. Soroa, R. Agerri, Applying deep learning techniques for sentiment analysis to assess sustainable transport, Sustainability 13 (4) (2021) http://dx.doi.org/10.3390/su13042397, URL https://www.mdpi.com/2071-1050/13/4/2397.

[25] I. Bouabdallaoui, F. Guerouate, S. Bouhaddour, C. Saadi, M. Sbihi, Named entity recognition applied on moroccan tourism corpus, Procedia Comput. Sci. 198 (2022) 373–378.

[26] X. Cheng, W. Wang, F. Bao, G. Gao, 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, Springer, 2020, pp. 11–23.

[27] C. Chantrapornchai, A. Tunsakul, Information extraction on tourism domain using SpaCy and BERT, ECTI Trans. Comput. Inf. Technol. 15 (1) (2021) 108–122.

[28] J.F. de Landa, R. Agerri, Social analysis of young Basque-speaking communities in twitter, J. Multiling. Multicult. Dev. (2021) 1–15.

[29] S. Kadam, V. Vaidya, 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, Springer, 2020, pp. 100–112.

[30] M. Moradi, K. Blagec, F. Haberl, M. Samwald, Gpt-3 models are poor few-shot learners in the biomedical domain, 2021, arXiv preprint arXiv:2109.02555.

[31] T. Schick, H. Schütze, 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, 2021, pp. 255–269.

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, CoRR, abs/1907.11692 arXiv:1907.11692 URL http://arxiv.org/abs/1907.11692.

[33] L. Wang, R. Li, Y. Yan, Y. Yan, S. Wang, W. Wu, W. Xu, InstructionNER: A multi-task instruction-based generative framework for few-shot NER, 2022, ArXiv Preprint, 2203.03903.

[34] I. García-Ferrero, R. Agerri, G. Rigau, Model and data transfer for cross-lingual sequence labelling in zero-resource settings, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, 2022, pp. 6403–6416.

[35] I. García-Ferrero, R. Agerri, G. Rigau, T-Projection: High quality annotation projection for sequence labeling tasks, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 15203–15217.

[36] A. Yeginbergen, R. Agerri, Cross-lingual argument mining in the medical domain, 2024, arXiv:2301.10527.

[37] M. Artetxe, G. Labaka, E. Agirre, Translation artifacts in cross-lingual transfer learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 7674–7684.

[38] M. Artetxe, V. Goswami, S. Bhosale, A. Fan, L. Zettlemoyer, Revisiting machine translation for cross-lingual classification, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 6489–6499.

[39] V. Barriere, A. Balahur, 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, International Committee on Computational Linguistics, 2020, pp. 266–271, http://dx.doi.org/10.18653/v1/2020.coling-main.23.

[40] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, K. Choukri, Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news, in: LREC, Citeseer, 2006, pp. 139–142.

[41] M. Taulé, M.A. Martí, M. Recasens, Ancora: Multilevel annotated corpora for Catalan and Spanish, in: Lrec, 2008.

[42] L. Derczynski, K. Bontcheva, I. Roberts, Broad Twitter corpus: A diverse named entity recognition resource, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 2016, pp. 1169–1179.

[43] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Proj. Rep. Stanf. 1 (12) (2009) 2009.

[44] H. Saif, M. Fernandez, Y. He, H. Alani, Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold, 2013.

[45] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, T. Wilson, 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), Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 312–320, URL https://aclanthology.org/S13-2052.

[46] S. Rosenthal, A. Ritter, P. Nakov, V. Stoyanov, SemEval-2014 task 9: Sentiment analysis in Twitter, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEVal 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 73–80, http://dx.doi.org/10.3115/v1/S14-2009, URL https://aclanthology.org/S14-2009.

[47] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, V. Stoyanov, SemEval-2016 task 4: Sentiment analysis in Twitter, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEVal-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 1–18, http://dx.doi.org/10.18653/v1/S16-1001, URL https://aclanthology.org/S16-1001.

[48] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 task 4: Sentiment analysis in Twitter, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEVal-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 502–518, http://dx.doi.org/10.18653/v1/S17-2088, URL https://aclanthology.org/S17-2088.

[49] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, M. Gašić, 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, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 5016–5026, http://dx.doi.org/10.18653/v1/D18-1547, URL https://aclanthology.org/D18-1547.

[50] S. Bowman, G. Angeli, C. Potts, C.D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 632–642.

[51] M. Masson, C. Sallaberry, R. Agerri, M.-N. Bessagnet, P. Roose, A. Le Parc Lacayrelle, 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, Springer, 2022, pp. 11–20.

[52] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266, URL https://aclanthology.org/2022.lrec-1.27.

[53] F. Barbieri, J. Camacho-Collados, L.E. Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1644–1650.

[54] J.M. Pérez, J.C. Giudici, F. Luque, Pysentimiento: A python toolkit for sentiment analysis and SocialNLP tasks, 2021, arXiv:2106.09462.

[55] Seethal, Sentiment analysis generic dataset, 2023, https://huggingface.co/Seethal/sentiment_analysis_generic_dataset. (Accessed 23 March 2023).

[56] J. Hartmann, M. Heitmann, C. Siebert, C. Schamp, More than a feeling: Accuracy and application of sentiment analysis, Int. J. Res. Mark. 40 (1) (2023) 75–87, http://dx.doi.org/10.1016/j.ijresmar.2022.05.005.

[57] L. Massaron, Fine-tune Mistral v0.2 for sentiment analysis — kaggle.com, 2024, https://www.kaggle.com/code/lucamassaron/fine-tune-mistral-v0-2-for-sentiment-analysis. (Accessed 20 April 2024).

[58] L. Massaron, Fine-tune Llama 2 for sentiment analysis — kaggle.com, 2024, https://www.kaggle.com/code/lucamassaron/fine-tune-llama-2-for-sentiment-analysis. (Accessed 20 April 2024).

[59] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2020, arXiv preprint arXiv:2009.03300.

[60] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6.

[61] E.F. Tjong Kim Sang, 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), 2002.

[62] M. Masson, C. Sallaberry, M.-N. Bessagnet, A.L.P. Lacayrelle, P. Roose, R. Agerri, 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, 2024, pp. 1–9.

[63] E.T. Hall, E.T. Hall, The Hidden Dimension, vol. 609, Anchor, 1966.