Information Technology & Tourism (2025) 27:227–261
https://doi.org/10.1007/s40558-024-00309-9

ORIGINAL RESEARCH

Discovering sentiment insights: streamlining tourism review analysis with Large Language Models

Dario Guidotti[1] · **Laura Pandolfo**[1] · **Luca Pulina**[1]

Received: 10 April 2024 / Revised: 24 December 2024 / Accepted: 27 December 2024 /
Published online: 17 January 2025

Abstract

With digital technology increasingly shaping the tourism industry, understanding customer sentiment and identifying key themes in reviews is crucial for enhancing service quality. However, traditional sentiment analysis and keyword extraction models typically demand significant time, computational resources, and labelled data for training. In this paper, we explore how Large Language Models (LLMs) can be leveraged to automatically classify reviews as positive or negative and extract relevant keywords without the need for dedicated training. Additionally, we frame the keyword extraction task as a tool to assist human users in comprehending and interpreting review content, especially in scenarios where ground truth labels for keywords are unavailable. To evaluate our approach, we conduct an experimental analysis using several datasets of tourism reviews and various LLMs. Our results demonstrate the reliability of LLMs as zero-shot classifiers for sentiment analysis and showcase the efficacy of the approach in extracting meaningful keywords from reviews, providing valuable insights into customer sentiments and preferences. Overall, this research contributes to the intersection of information technology and tourism by presenting a practical solution for sentiment analysis and keyword extraction in tourism reviews, leveraging the capabilities of LLMs as versatile tools for enhancing decision-making processes in the tourism industry.

Keywords AI for tourism · Sentiment analysis · Natural language processing · Keyword extraction

✉ Dario Guidotti
dguidotti@uniss.it

Laura Pandolfo
lpandolfo@uniss.it

Luca Pulina
lpulina@uniss.it

[1] Artificial Intelligence and Formal Methods Laboratory, Department of Humanities and Social Sciences, University of Sassari, Via Roma 151, 07100 Sassari, Sardinia, Italy

1 Introduction

The tourism industry, like many other sectors, is experiencing profound shifts due to the pervasive integration of digital technology into its operations (Pencarelli 2020). This ongoing evolution has prompted a critical imperative for tourism businesses: the ability to effectively interpret and respond to the sentiments conveyed within customer reviews. As travellers increasingly rely on digital platforms to share their experiences, the wealth of information contained within these reviews presents a valuable opportunity for businesses to gain insights into customer preferences, satisfaction levels, and areas for improvement (Rossetti et al. 2016; Jardim and Mora 2021).

Sentiment analysis and keyword extraction stand as two Natural language processing (NLP) tasks that provide these capabilities. Specifically, sentiment analysis facilitates the automatic classification of texts (e.g., reviews, social media comments) into positive, negative, or neutral categories, while keyword extraction enables the distillation of primary themes and topics expressed within such texts. However, traditional methods of sentiment analysis and keyword extraction present considerable obstacles in this endeavour. These approaches often demand significant investments of time, computational resources, and access to labelled data for training (Sharma et al. 2022). Such requirements can prove prohibitive, particularly for smaller businesses or those operating within niche markets where data availability may be limited. Consequently, businesses face challenges in harnessing the full potential of customer feedback to inform strategic decision-making and enhance service quality. Zero-shot learning (ZSL) emerges as a potential solution to overcome these limitations: this methodology capitalises on semantic embeddings to deduce insights about unseen data based on their correlation with known data. Consequently, these models can be directly applied without the necessity for task-specific training.

In this paper, we showcase the potential of Large Language Models (LLMs) to streamline sentiment analysis and keyword extraction in tourism reviews. Unlike conventional approaches, which often rely on laborious processes involving the manual labelling of vast datasets for training, these models harnesses the power of zero-shot learning to automatically classify reviews as positive or negative and extract pertinent keywords without the need for specialised training data. To assess the effectiveness of our approach, we conduct an experimental analysis evaluating multiple datasets of tourism reviews and various LLM architectures. The findings of our study assess the reliability of LLMs as zero-shot classifiers for sentiment analysis and underscore the efficacy of our approach in extracting meaningful keywords from reviews. These insights offer valuable perspectives into customer sentiments and preferences, thereby providing tourism Small Medium Enterprises (SMEs) with actionable intelligence for enhancing decision-making processes. To the extent of out knowledge, the presented study is the first to explore the utilisation of zero-shot learning models for our tasks of interest or to conduct a comparison between such models.

In summary, this study contributes to the intersection of digital technology and tourism by presenting a practical solution for sentiment analysis and keyword extraction in tourism reviews. By harnessing the versatility of LLMs, we lay the groundwork for informed decision-making and enhanced service delivery within the ever-evolving tourism sector. Furthermore, by emphasising open-source LLM architectures that do not necessitate specialised training from end-users, the proposed solution can be readily utilised, even by SMEs that may struggle to acquire the financial and computational resources required to train NLP models independently. In the framework of the proposed work, one of our objectives is to prioritise technology transfer to actively involve local communities in digital innovation, thus propelling towards a knowledge-based economy. This involves supporting SMEs, which are crucial components of our regional economy, in the adoption of new technologies to bolster their current and future competitiveness.

The rest of the paper is structured as follows. In Sect. 2 we introduce some basic concepts and definitions, while in Sect. 3 we present the related work. In Sect. 4 we describe the presented approach and the experimental setup. Finally, in Sects. 5 and 6 we present the results of our experimental evaluation and summarise our conclusions, respectively.

2 Background

2.1 Sentiment analysis and keyword extraction

Sentiment analysis is a NLP task aimed at determining the sentiment expressed in a piece of text. With the proliferation of user-generated content on digital platforms, sentiment analysis has emerged as a vital tool for businesses to understand and respond to customer opinions, feedback, and emotions. By automatically classifying text as positive, negative, or neutral, sentiment analysis enables organisations to gauge customer satisfaction, identify areas for improvement, and tailor their products or services to meet customer expectations. Traditional approaches to sentiment analysis often rely on Machine Learning (ML) algorithms trained on labelled datasets, where each text sample is annotated with its corresponding sentiment polarity. These algorithms learn to classify new text samples based on patterns and features extracted from the training data. While effective, this supervised learning paradigm requires substantial amounts of labelled training data and may struggle to generalise to new domains or languages where labelled data is scarce. For further exploration of the various approaches, we refer to Birjali et al. (2021) and Wankhade et al. (2022).

Keyword extraction, on the other hand, involves the automatic identification of significant words or phrases that capture the essence of a document or text corpus. Keywords play a crucial role in summarising the main themes, topics, or sentiments expressed within textual content, facilitating information retrieval, analysis, and interpretation. Traditional keyword extraction methods often rely on statistical metrics, such as term frequency-inverse document frequency (TF-IDF) or graph-based algorithms like TextRank, to identify important keywords based on their frequency,

importance, or connectivity within a document (Firoozeh et al. 2020). However, traditional keyword extraction methods may face challenges in handling noisy or unstructured text data, as well as in identifying contextually relevant keywords that capture the nuances of the text. Moreover, the effectiveness of keyword extraction approaches may be limited by the availability of labelled training data or the need for manual parameter tuning.

In recent years, advancements in deep learning and the availability of large-scale pre-trained language models, have revolutionised the field of sentiment analysis and keyword extraction.

2.2 Large Language Models

Large Language Models stand at the forefront of contemporary natural language processing, embodying a substantial leap in the field's capabilities. These models, rooted in transformer architectures, have emerged as revolutionary tools for comprehending and generating human-like text across a diverse array of tasks and domains (Abdullah and Ahmet 2023). Transformers, introduced by Vaswani et al. (2017), depart from earlier sequence-to-sequence models by dispensing with Recurrent Neural Networks (RNNs) in favour of self-attention mechanisms. This mechanism allows the model to weigh the importance of each token in the input sequence, facilitating the capture of long-range dependencies and contextual relationships within the text.

One of the defining features of LLMs is their extensive pre-training on a vast corpora of text data. During pre-training, the model learns to predict missing words or tokens within text sequences, leveraging the wealth of linguistic patterns and structures present in the data. This pre-training phase instils LLMs with a rich understanding of language semantics, syntax, and context, enabling them to generalise across a wide spectrum of language tasks and domains. Following pre-training, LLMs can be fine-tuned on task-specific datasets to adapt to particular applications or domains. Fine-tuning involves updating the model's parameters using labelled data from the target task, thereby tailoring the model to excel in specific linguistic tasks such as sentiment analysis, text classification, or language generation. Furthermore, LLMs support transfer learning, where knowledge acquired during pre-training can be transferred to downstream tasks with minimal task-specific training data, enhancing their versatility and applicability.

LLMs have garnered widespread acclaim for their scalability and performance, with state-of-the-art models such as BERT (Devlin et al. 2019) and GPTs (Generative Pre-trained Transformers) comprising billions of parameters (Wang et al. 2022). These large-scale models exhibit remarkable capabilities in generating coherent and contextually relevant text, surpassing previous benchmarks on a myriad of language understanding and generation tasks. In the realm of sentiment analysis and keyword extraction, LLMs offer unparalleled advantages. Their deep understanding of language semantics and context enables them to accurately infer sentiment polarity and identify relevant keywords within textual data. Moreover, the zero-shot learning capabilities of LLMs allow them to perform these tasks

without explicit training on labelled datasets, obviating the need for extensive task-specific training. For a more detailed overview regarding LLMs we refer to Raiaan et al. (2024).

2.3 Zero-shot learning

Zero-shot learning stands as a transformative approach within the realm of machine learning, revolutionising the traditional paradigm of supervised learning by enabling models to generalise to unseen classes or tasks without requiring explicit training on labelled examples (Xian et al. 2019; Wang et al. 2019). Unlike conventional supervised methods, which rely heavily on annotated data for each class or task of interest, ZSL leverages auxiliary information or semantic embedding to infer knowledge about unseen classes based on their relationships to known classes. Initially conceived within the domain of computer vision to address challenges in image classification, ZSL has since expanded its applicability to various fields, including NLP. In this area, the principles of ZSL have opened new avenues for models to perform tasks such as text classification, sentiment analysis, and language generation without the need for labelled training data specific to each task.

The foundation of ZSL lies in its utilisation of semantic representations to bridge the gap between known and unseen classes or tasks. These semantic embedding, which may include word embedding, language embedding, or ontological knowledge graphs, encode the semantic relationships between different classes or concepts within a continuous vector space (Wang et al. 2018). By leveraging these representations, ZSL models can generalise across tasks or domains, transferring knowledge from known classes to unseen ones. A key advantage of ZSL is its reliance on transfer learning principles, where knowledge acquired from pre-training on large-scale datasets or auxiliary tasks can be transferred to related tasks or domains. This enables ZSL models to acquire generalised knowledge about language semantics, syntax, and context, which can then be adapted to perform specific tasks with minimal task-specific training data. Furthermore, ZSL often involves task decomposition techniques, where complex tasks are broken down into simpler sub-tasks or components. By learning to perform these sub-tasks independently, models can generalise to unseen tasks by combining their knowledge of known sub-tasks in novel ways, thereby enhancing their adaptability and versatility. A particularly relevant task is Natural Language Inference (NLI), which involves determining whether the information presented in one text can be inferred from another text. This task can be utilised to develop models capable of categorising text into generic classes as specified by users, without the necessity for training or fine-tuning. Further details on NLI are provided in the subsequent section.

In recent years, advancements in ZSL have led to the development of sophisticated models capable of zero-shot text classification, sentiment analysis, and keyword extraction. These models leverage semantic embedding and transfer learning techniques to generalise across tasks and domains, enabling them to perform effectively in scenarios where labelled training data is scarce or unavailable.

2.4 Natural language inference

Natural language inference, also known as Textual Entailment, is a central task in natural language processing that focuses on determining whether a given text, referred to as the premise, logically entails another text, known as the hypothesis. This task plays a crucial role in assessing a model's ability to understand and reason about language, which is fundamental for a wide range of NLP applications such as question answering, summarisation, and information retrieval. It is important to note that textual entailment differs from classical logical entailment, as it has a more relaxed definition: a text t entails an hypothesis h if a human reading t would infer that h is most likely true (Dagan et al. 2005). In traditional NLI tasks, the relationship between the premise and the hypothesis is classified into three categories: entailment, contradiction, and neutral. If the hypothesis logically follows from the premise, it is classified as entailment. If the hypothesis is inconsistent with the premise, it is classified as contradiction. If there is no significant relationship between the premise and the hypothesis, it is classified as neutral. In this work, we focus on models that leverage NLI for zero-shot text classification. These models treat the classes as hypotheses: if the model determines that the text entails a hypothesis, the text is classified under that category. Conversely, if the model finds a contradiction or neutrality between the text and the hypothesis, it indicates that the text does not belong to that category. Despite significant progress, NLI remains challenging due to the inherent ambiguity and vagueness of natural language. Many entailment decisions require commonsense knowledge or external world information, which current models may lack. Multi-hop reasoning, where a hypothesis requires integrating information from multiple premises, adds another layer of complexity to the task. However the introduction of transformer models marked a significant breakthrough, setting new benchmarks in NLI by leveraging pre-training on large corpora followed by fine-tuning on NLI datasets.

3 Related work

In the current state-of-the-art, sentiment analysis classification approaches are typically divided into lexicon-based and ML-based methods.

Lexicon-based approaches utilise sentiment lexicons, which consist of opinion words and phrases annotated with positive or negative scores, to determine the sentiment polarity of terms in text (Bagherzadeh et al. 2021; Bucur 2015; Gräbner et al. 2012; Hnin et al. 2018). Their primary advantage lies in their independence from labelled data. However, they rely heavily on linguistic resources and struggle to account for context. In this study, we do not further focus on this kind of methodologies due to the scope of our current investigation. For a more thorough examination of these approaches, we recommend referring to Ameur et al. (2024) for additional insights and details.

ML-based methods do not necessitate a predefined dictionary and demonstrate greater proficiency in handling ambiguities and adapting to different domains. However, training such models is often time-consuming and typically relies on

labelled data. These approaches can be further categorised into shallow learning methods and deep learning methods.

Some of the main shallow learning methods used in sentiment analysis are Naive Bayes (NB) and Support Vector Machines (SVMs). NB-supervised classification utilises Bayes' rule to assign the most probable class to a certain text. Notably, NB exhibits acceptable precision and low computational cost, as noted by Martins et al. (2017), and is robust against noise and overfitting, making it a popular choice for sentiment classification in hotel reviews (Farisi et al. 2019; Ghorpade and Ragha 2012; Martins et al. 2017). SVMs, unlike probabilistic classifiers, work by finding the optimal hyperplane that best separates different classes of data. Such hyperplane is positioned to maximise the margin between the closest data points of each class, resulting in a robust and efficient classification model. Studies suggest that SVM outperforms other shallow ML algorithms in accuracy for sentiment analysis (Shi and Li 2011).

Deep learning methodologies involve complex neural network architectures with multiple hidden layers, enabling them to autonomously learn and refine their own representations of the data. The primary deep learning models employed in sentiment analysis tasks comprise Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer language models. Jiang et al. (2019) applied CNN layers to identify local features for aspect sentiment classification in Chinese hotel reviews, while de Souza et al. (2018) employed CNNs for classifying Brazilian hotel reviews, yielding promising results. However, while CNNs are proficient in analysing short sentences, RNNs may be more suitable for longer ones (de Souza et al. 2018). Long Short-Term Memory (LSTM), a popular RNN architecture, maintains information in memory for extended periods, offering exceptional performance in sentiment analysis (Pal et al. 2018). Priyantina and Sarno (2019) combined word embedding with LSTM for sentiment classification of reviews, achieving high accuracy for aspect-based sentiment analysis. Despite LSTM architectures' capability to capture contextual information, they struggle with identifying crucial corpus parts (Jiang et al. 2019). Younas et al. (2020) examined the performance of two transformer models, multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R), for sentiment analysis of multilingual social media text. Tan et al. (2022) introduced an advanced sentiment analysis approach by merging the strengths of two prominent models: the robustly optimised BERT approach (RoBERTa) and LSTM. Finally, Tesfagergish et al. (2022) proposed a two-stage emotion detection methodology for sentiment analysis leveraging zero-shot learning models and trained ensembles. Evaluation across three benchmark datasets yielded the best accuracy of 87% for binary sentiment classification and 63% for the three-class sentiment classification.

While our methodology shares some similarities with this last work, it is important to highlight that Tesfagergish et al. (2022) trains the ensembles from scratch, whereas we directly employ zero-shot learning models without requiring fine-tuning or specifically trained models. For a more detailed survey of sentiment analysis methodologies we refer to Ameur et al. (2024).

4 Materials and methods

In the following, we introduce the models and datasets examined in our experimental evaluation, offering insight into the rationale behind our selections. Additionally, we focus on the proposed methodologies for sentiment analysis and keyword extraction tasks with greater details.

4.1 Models

In our experimental evaluation, we used a selection of state-of-the-art NLP zero-shot classifiers, readily available from the Hugging Face model hub,[1] a well-known platform for NLP model deployment and experimentation. These models were chosen for their accessibility and their capabilities specifically tailored for zero-shot classification tasks, which allow them to generalise effectively to unseen classes or categories without the need for task-specific training data. Moreover, we prioritise models sourced from reputable origins, preferably with accompanying references, and not tailored for languages other than English. In addition to the zero-shot models, we also considered a selection of four State-Of-The-Art (SOTA) models optimised for sentiment analysis. This approach aims to elucidate the trade-off between the accuracy of models specifically trained for sentiment analysis and the versatility of the zero-shot classifiers, which can be used for both sentiment analysis and keyword extraction tasks. For clarity, we refer to the zero-shot classifiers as $M1$-$M9$ and to the SOTA models as $S1$-$S4$. It should be noted that $S4$ is fine-tuned to recognise only positive and negative sentiment polarity, whereas all the other models can be used to also recognize the neutral polarity.

The models shown in Table 1 offer varying degrees of complexity, from large-scale architectures capable of handling extensive text data to smaller, more lightweight models suitable for less resource-intensive tasks. As can be seen, many of them originate from the same base models: this is due to the limited availability of open source LLMs which consequently constrains the variety of their derived models. However, it should be noted that the training process is non-deterministic—meaning, two models trained on the same datasets may exhibit different behaviours. Additionally, smaller models are not necessarily inferior to larger ones across various tasks of interest. Specifically, when evaluating a single review, the response of the most complex model may not always be more reliable than that of the smallest model. Furthermore, it is essential to consider that larger models require correspondingly higher computational resources for inference execution. For these latter reasons, we have chosen to include all the models presented, even those belonging to the same "family" and that present decreasing levels of complexity. In particular, we considered models originating from four different LLMs:

[1] https://huggingface.co.

Table 1 Models considered in our experimental evaluation

Mod ID	Description
M1 [a]	A model by Laurer et al. (2023) based on the large version of the DeBERTa model (He et al. 2021) and fine-tuned on several general classification dataset
M2 [b]	A model based on the base version of the DeBERTa model and fine-tuned on several general classification dataset
M3 [c]	A model based on the extra small version of the DeBERTa model and fine-tuned on several general classification dataset
M4 [d]	A model based on the large version of the BART model (Lewis et al. 2020) and fine-tuned on the MNLI dataset (Williams et al. 2018)
M5 [e]	A smaller model based on MiniLM2 (Wang et al. 2021) and trained on the SNLI (Bowman et al. 2015) and MNLI datasets
M6 [f]	A model based on the large version of the DeBERTa model and trained on the SNLI and MNLI datasets
M7 [g]	A model based on the extra small version of the DeBERTa model and trained on the SNLI and MNLI datasets
M8 [h]	A model based on the small version of the DeBERTa model and trained on the SNLI and MNLI datasets
M9 [i]	A model based on the base version of the DeBERTa model and trained on the SNLI and MNLI datasets
S1 [j]	A model (Loureiro et al. 2022) based on the base version of the RoBERTa (Liu et al. 2019) model and fine-tuned for sentiment analysis with the TweetEval benchmark (Barbieri et al. 2020)
S2 [k]	A multilingual model (Barbieri et al. 2021) based on the base version of the RoBERTa model and fine-tuned for sentiment analysis
S3 [l]	A model based on a specialised version (Nguyen et al. 2020) of the RoBERTa model and fine-tuned on the SemEval-2017 dataset (Rosenthal et al. 2017)
S4 [m]	A model (Hartmann et al. 2023) based on the large version of the RoBERTa model and fine-tuned for sentiment analysis on 15 datasets

Column Mod ID represent the identifiers used for the models. Column Description provides a brief description of the corresponding models

[a] https://huggingface.co/MoritzLaurer/deberta-v3-large-zeroshot-v1

[b] https://huggingface.co/MoritzLaurer/deberta-v3-base-zeroshot-v1

[c] https://huggingface.co/MoritzLaurer/DeBERTa-v3-xsmall-mnli-fever-anli-ling-binary

[d] https://huggingface.co/facebook/bart-large-mnli

[e] https://huggingface.co/cross-encoder/nli-MiniLM2-L6-H768

[f] https://huggingface.co/cross-encoder/nli-deberta-v3-large

[g] https://huggingface.co/cross-encoder/nli-deberta-v3-xsmall

[h] https://huggingface.co/cross-encoder/nli-deberta-v3-small

[i] https://huggingface.co/cross-encoder/nli-deberta-v3-base

[j] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[k] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[l] https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis

[m] https://huggingface.co/siebert/sentiment-roberta-large-english

- **BART** (Lewis et al. 2020) is designed as a versatile sequence-to-sequence model tailored for tasks such as text generation, summarisation, translation, and question answering. It combines features from both bidirectional transformers (like BERT) and autoregressive transformers (like GPT), allowing it to excel in a wide range of NLP applications.
- **DeBERTa** (He et al. 2021) builds upon BERT by introducing a novel disentangled attention mechanism, where each word is represented by one vector for content and one for position. This enhances the model's ability to capture intricate relationships between words within a sentence. These advancements contribute to DeBERTa achieving state-of-the-art performance across various NLP benchmarks.
- **MiniLM2** (Wang et al. 2021) is engineered to be a more compact and efficient alternative to larger language models like BERT, while still delivering competitive performance. It achieves this through a process of knowledge distillation from larger models, which allows it to retain much of the performance benefits while reducing computational demands. Despite its smaller size, MiniLM2 demonstrates strong performance across a range of NLP tasks, making it suitable for deployment in resource-constrained environments.
- **RoBERTa** (Liu et al. 2019) builds upon BERT (Devlin et al. 2019) by optimising its pre-training procedure. This includes training on more data, longer sequences, and removing the next sentence prediction task. These enhancements lead to improved language understanding capabilities and robust performance across multiple NLP benchmarks.

By leveraging this diverse set of pre-trained NLP models, we aimed to comprehensively evaluate their performance in the context of sentiment analysis and keyword extraction tasks within tourism domain reviews. This approach enabled us to assess the effectiveness and robustness of zero-shot learning techniques across a range of model architectures, providing valuable insights into their suitability for real-world applications in the tourism industry. Furthermore, by considering together the responses of the separate models as an ensemble, we are able to gather more insight both on their performances and on the reviews considered.

4.2 Datasets

In our experimental evaluation, we selected diverse datasets to ensure a comprehensive assessment of the proposed methodology across different contexts. The datasets chosen for analysis include:

1. **European Castles Dataset**.[2] This dataset comprises 2550 Google reviews encompassing 529 European castles. These reviews provide useful insights into tourists' experiences and perceptions of historical landmarks, allowing us to explore

[2] https://www.kaggle.com/datasets/datasciencedonut/european-castles.

Fig. 1 Histograms of the distributions of reviews for each possible score and dataset

sentiment analysis and keyword extraction within the realm of cultural heritage tourism. We will use the identifier **Castles** to refer to this dataset from now on.

2. **Google Maps Restaurant Reviews Dataset**.[3] With 1100 Google reviews spanning 100 different restaurants, this dataset provides a rich source of information on dining experiences and culinary preferences. Analysing sentiment and extracting keywords from restaurant reviews enables us to focus on the gastronomic and culinary tourism domains. We will use the identifier **Restaurants** to refer to this dataset.
3. **Hotel Reviews Dataset**.[4] Consisting of 10000 Google reviews pertaining to 1000 hotels across the United States, this dataset offers an overview of accommodation experiences in the hospitality sector. By examining sentiment and identifying keywords within hotel reviews, we collect insights into travellers' preferences, satisfaction levels, and expectations within the lodging industry. We will use the identifier **Hotels** to refer to this dataset.

In Fig. 1, we provide a concise summary of the distribution of reviews within each dataset across different rating scores. It is evident that the datasets we examined display relevant imbalances concerning the quantity of positive (greater than three stars), neutral (three stars), and negative (less than three stars) reviews. Recognising these imbalances is crucial for accurately interpreting the findings derived from our experimental evaluation. Furthermore, it is important to note that such disparities are typical in datasets encompassing reviews even across different domains.

In Fig. 2, we present a graphical representation of the distribution of reviews based on their text length. The distributions of the datasets are quite similar, yet they are centred around different lengths. In particular, the **Restaurants** dataset predominantly features shorter reviews, whereas the **Castles** and **Hotels** datasets tend to contain significantly longer reviews. This observation aligns with the nature of the respective domains: there is generally more to discuss regarding a hotel stay or a visit to a cultural heritage site than a single dining experience.

[3] https://www.kaggle.com/datasets/denizbilginn/google-maps-restaurant-reviews.
[4] https://www.kaggle.com/datasets/datafiniti/hotel-reviews.

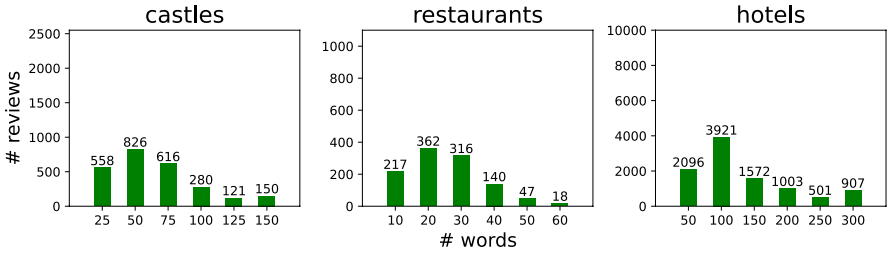

Fig. 2 Histograms of the distributions of reviews with respect to their length for each dataset. Each bin reports the number of reviews with length less or equal to the corresponding tick. For the sake of clarity, the last bin also includes reviews with length greater than the corresponding bin

By leveraging these diverse datasets, we aim to evaluate the effectiveness and robustness of the proposed methodology across different tourism domains. The inclusion of datasets spanning castles, restaurants, and hotels allows for a holistic assessment of sentiment analysis and keyword extraction techniques, providing valuable insights applicable to a wide range of tourism-related businesses and destinations. Moreover, it is noteworthy that the language and communication styles employed in reviews may exhibit substantial variation across these datasets, thereby augmenting the reliability and applicability of our findings to diverse contexts. Finally, the significant variation in review lengths further contributes to the breadth of our experimental evaluation.

While our focus on dataset of reviews, which typically include a related number of stars, may make sentiment analysis appear superfluous, it remains relevant. The reliability of scores is generally lower than the sentiment expressed in the textual portion of the reviews due to potential errors or misunderstandings by the writer. Therefore, sentiment analysis is essential for accurately measuring the true sentiment conveyed in the text. Furthermore, while we did not consider at this time dataset originating from social networks, we expect the results obtained in this work to directly apply also to social media texts related to particular tourist attractions. Note that the datasets used for this work do not include ground truth annotations specifically tailored for keyword extraction tasks. Currently, there is a notable absence of datasets designed specifically for keyword extraction within the tourism domain. Existing datasets typically originate from specialised fields [*e.g.*, extracting keywords from scientific literature (Kim et al. 2010; Augenstein et al. 2017)] and are limited in terms of manually annotated texts. This lack of domain-specific datasets underscores the rationale behind our approach of leveraging LLMs for automated keyword extraction. Establishing the reliability of this methodology could facilitate automatic annotation for future datasets in tourism research.

4.3 Methodology

In this subsection, we offer a comprehensive overview of the methodology proposed in our study for conducting sentiment analysis and keyword extraction tasks within

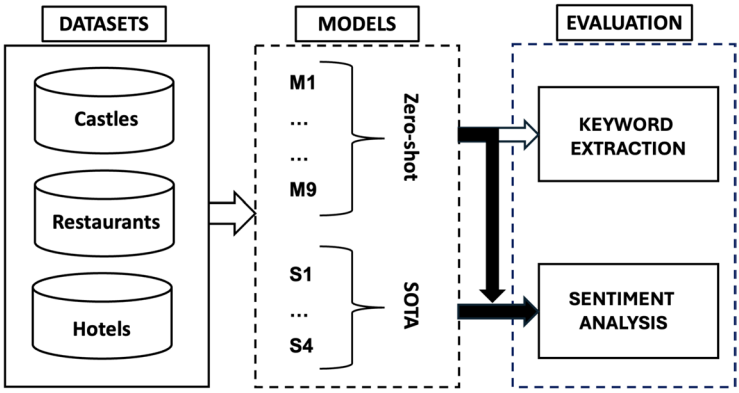

Fig. 3 Workflow diagram illustrating the experimental setup for sentiment analysis and keyword extraction across Castles, Restaurants, and Hotels datasets using nine zero-shot and four state-of-the-art models

tourism reviews. For sentiment analysis, we classify reviews as positive, neutral, or negative based on their provided scores, aiming to assess the accuracy of NLP models in discerning sentiment. In keyword extraction, we preselect five keywords of interest per dataset and employ NLP models to determine the probabilities of each keyword to be relevant for each review. We then evaluate the results using statistical analysis.

In Fig. 3 we present the workflow diagram of our experimental evaluation. It provides an overview of the experimental framework applied to the three distinct datasets. It should be noted that the inputs of our models consist of reviews expressed as raw texts, without other specific prompts. In our analysis, we incorporate two categories of models: nine zero-shot models and four SOTA models. For sentiment analysis, we employ all models, both zero-shot and SOTA, to assess the sentiment conveyed in the textual data across each dataset. In contrast, keyword extraction used only the nine zero-shot models to identify and extract significant terms from the textual content within each dataset.

4.3.1 Sentiment analysis

In our study, the sentiment analysis task revolves around the classification of reviews based on their sentiment polarity. To achieve this, we employ the transformer pipeline,[5] which is a user-friendly library for inference and fine-tuning using transformers models provided by Hugging Face. Utilising Python scripts, we interface with the transformer pipeline to apply the models of interest to every review within the datasets. For the NLI based models ($M1$-$M9$), we provide them with *positive*, *negative*, or *neutral* as three mutually exclusive hypothesis. Therefore the actual output of these models consists in three score between 0 and 1 corresponding to the likelihood

[5] https://huggingface.co/docs/transformers/index.

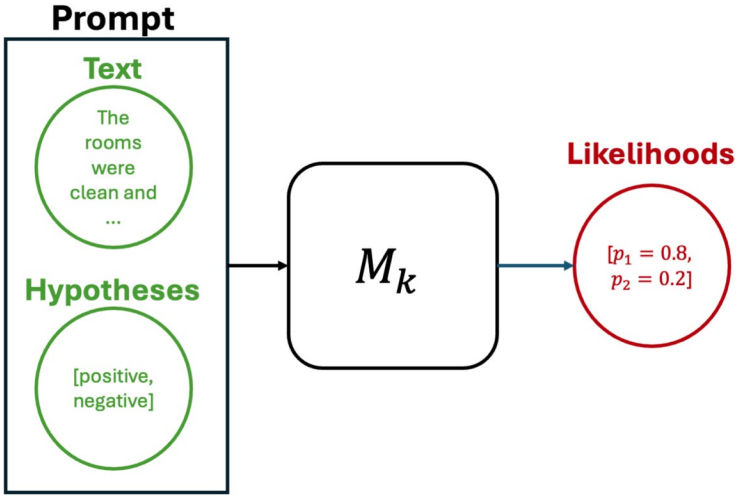

Fig. 4 Illustration of the NLI model's input-output process. The prompt consists of the review text and a set of possible hypotheses. The model processes this input to produce likelihoods for each hypothesis, which are mutually exclusive and sum to 1

that the input text entails the specific hypothesis. As the classes are mutually exclusive, the sum of the three scores is equal to 1. To further clarify, for all the NLI models considered, the prompt consisted of the text of the review under consideration and a list of possible hypotheses. The model's output was a set of mutually exclusive likelihoods $p_i \in [0, 1] \subset \mathbb{R}$, such that $\sum_i p_i = 1$, representing the probabilities that the review corresponded to each specific class. For a visual representation of the relationship between the *prompt*, the *model*, and the corresponding *output*, please refer to Fig. 4.

The SOTA models ($S1$–$S4$) are fine-tuned for sentiment analysis, therefore their output already consists by default in the likelihood that the input text could be associated to the sentiment label of interest. However, $S4$ supports only the positive and negative classes, and therefore has been excluded from the evaluation when the neutral sentiment is considered. The response of the model is then determined as the sentiment label (positive, neutral, or negative) associated with the highest likelihood. Then this response is compared with the ground truth to provide an estimation of the accuracy of the models on the task of interest. After this initial experimental evaluation, we opted to expand our analysis to include scenarios where the reviews presenting a score of three stars, corresponding to a neutral polarity, were removed from the dataset.

In addition to assessing the accuracy of the models, we compute standard NLP measurements of *Precision*, *Recall*, and *F Score*. *Precision* refers to the proportion of correctly classified reviews out of all reviews classified as a certain class by the model. *Recall*, on the other hand, represents the proportion of correctly identified reviews out of all the reviews actually belonging to the class of interest in the dataset. Finally, *F Score* is the harmonic mean of *Precision* and *Recall*, providing a balanced measure of the model's performance. We compute these metrics to gain a

more comprehensive understanding of the models' effectiveness in sentiment analysis. While accuracy provides a general overview, *Precision*, *Recall*, and *F Score* offer more insights into the models' ability to correctly identify the sentiment polarity of the reviews of interest. Similar to accuracy, these three metrics range between 0 and 1, where a higher value indicates better performance.

As our ground truth for sentiment classification, we adopt a straightforward criterion: a review is considered positive if its score is greater than 4 stars, neutral if its score is 3 stars, and negative otherwise.

4.3.2 Keywords extraction

The first step in our methodology for identifying relevant keywords in specific reviews is to define a set of target keywords. This approach is necessary because using LLMs directly for keyword extraction may lead to the detection of uninformative terms (*e.g.*, "food" for restaurant reviews, "room" for hotel reviews, etc.). In our experiments, the target keywords were selected through a systematic process involving collaboration with domain experts in the tourism sector as part of a technology transfer initiative aimed at supporting SMEs. Initial steps involved conducting targeted surveys among a representative sample of 50 SMEs from the tourism and cultural heritage sectors. These surveys were designed to gather insights into the keywords that SMEs themselves identified as most relevant and impactful for their domains. To ensure a comprehensive experimental evaluation, two different sets of keywords were selected: the first set was chosen without giving any aid to the experts and the SMEs, whereas for the second set they were aided with the word clouds presented in Fig. 5, thus incorporating a data-driven perspective. This dual approach balances expert intuition with empirical evidence. It is important to note that although the word clouds were presented to the SMEs, they were not restricted to choosing only the keywords contained within them. For instance, they selected the term "price" for the **Castles** dataset upon observing that "ticket" was present in the corresponding word cloud, as they deemed it to be more informative within the same thematic context.

Table 2 presents the sets of keywords chosen for each dataset: *SET1* denotes the keywords selected by domain experts without additional aid, while *SET2* represents those chosen with the assistance of word clouds presented in Fig. 5. It is important to note that the algorithm generating the word cloud already eliminates less informative terms such as articles and pronouns. As can be seen, relying solely on NLP for direct keyword selection may lead to sub-optimal choices: the word clouds might suggest "Hotel", "Room", and "Thank" as prominent keywords for the **Hotel** Dataset; "Good", "Place", and "Taste" for the **Restaurant** Dataset; "Castle", "Place", "Inside" for the **Castle** Dataset. Clearly, these keywords do not provide useful information to the end user. As anticipated, there is a degree of overlap between the two sets of keywords chosen for each dataset.

Once the set of keywords of interest is identified, the models are employed to assess every review in our datasets, determining the likelihood of each keyword in the sets being associated with each review. To accomplish this, we provide to each model the selected keywords as not mutually exclusive hypotheses. Therefore their outputs consist

Table 2 Keyword Sets considered in our experimental evaluation

Dataset	Key set	Keywords
Castles	*SET 1*	Tour, History, Panorama, Staff, Food
	SET 2	Tour, History, View, Garden, Price
Restaurants	*SET 1*	Price, Staff, Clean, Comfortable, Location
	SET 2	Price, Service, Pizza, Meat, Breakfast
Hotels	*SET 1*	Price, Staff, Location, Food, Ambience
	SET 2	Price, Staff, Location, Breakfast, Clean

Column Dataset represent the dataset of interest. Column Key Set reports the identifiers of the keyword sets. Note that SET1 denotes the keywords selected by domain experts without additional aid, while SET2 represents those chosen with the assistance of word clouds. Finally, Column Keywords reports the keywords corresponding to each set

Fig. 5 Word Clouds generated for the Datasets of interest. Larger font sizes correspond to higher frequencies of the respective words in the dataset. Different font colours are utilised for readability purposes, with no additional information conveyed

in the scores, between 0 and 1, corresponding to the likelihood that the input text entails each specific hypothesis, and more than one hypothesis, or even none, can be entailed at the same time. The prompt schema is the same as the one in Fig. 4 with the only difference that the hypotheses of interest are our selected keywords and the likelihoods are not mutually exclusive.

It is important to highlight that, unlike the sentiment analysis task, we lack a ground truth in this scenario. Therefore, once we obtain the probabilities provided by each model for each review and keyword, we conduct statistical analysis to analyse the distribution of these probabilities. Specifically, we compute the median and interquartile (IQ) range of the probabilities provided by the models for each keyword in each review, and we analyse the distribution of these measurements over the whole dataset. Considering two sets of keywords also allows us to gather more insight in the behaviour of the models. This analysis enables us to gain insights into the consistency and variability of model responses, providing a comprehensive understanding of the models' performance in keyword extraction. Additionally, we analyse the distribution of probabilities for each keyword considering selected reviews to better understand the differences between the responses of various models. This targeted analysis allows us to identify potential patterns or discrepancies in model behaviour and performance.

5 Results and discussion

In this section, we present and discuss the outcomes of our experimental assessment for sentiment analysis and keyword extraction. Furthermore, we offer hypotheses and considerations concerning our results, exploring specific reviews for thorough analysis.

5.1 Sentiment analysis

In this sub-section we present our results considering three potential polarities for the reviews: positive, neutral, and negative. As depicted in Table 3, all the models under consideration struggle to accurately identify reviews with a neutral polarity, with the best case being model *M2* with the **Castles** dataset reaching 34.5% of accuracy. Conversely, they demonstrate relative proficiency in identifying positive reviews, albeit slightly less so in the case of negative reviews. In particular, in the case of the **Castle** dataset, the models exhibit greater success in identifying negative reviews compared to positive ones, contrasting with results from other datasets. Regarding the performance of the various models, no clear hierarchy in accuracy emerges, except for models *M6* and *M8* and *S2*, which appear to exhibit lower average accuracy compared to others. It is also interesting to note that the SOTA models do not seem to be significantly more accurate than the zero-shot models, even if their performances seem more consistent across different datasets. *S4* has not be considered in this first evaluation as it does not support the neutral polarity.

Further insight can be obtained by analysing Table 4. Indeed, the models appear to be proficient in recognising positive polarity, as evidenced by the presented

Table 3 Results of the sentiment analysis task when considering reviews with a score of three stars as neutral

Mod ID	Castles				Restaurants				Hotels			
	Acc	*Pos*	*Neut*	*Neg*	*Acc*	*Pos*	*Neut*	*Neg*	*Acc*	*Pos*	*Neut*	*Neg*
M1	90.2%	92.5%	19.5%	98.8%	75.6%	86.7%	12.8%	90.1%	84.8%	94.8%	9.8%	96.3%
M2	88.0%	89.7%	34.5%	96.3%	73.4%	81.2%	27.3%	85.5%	84.2%	93.0%	16.6%	95.4%
M3	88.7%	91.6%	10.3%	88.8%	76.3%	90.3%	11.0%	78.3%	81.2%	92.3%	5.0%	85.4%
M4	89.7%	92.6%	5.7%	95.0%	75.5%	89.2%	5.2%	85.5%	82.7%	93.3%	3.3%	93.7%
M5	90.9%	93.9%	6.9%	93.8%	74.6%	90.2%	7.0%	71.7%	81.8%	94.0%	7.3%	77.7%
M6	82.0%	84.1%	13.8%	91.3%	71.7%	84.5%	9.9%	76.3%	77.3%	86.9%	10.3%	82.3%
M7	90.4%	93.4%	9.2%	91.3%	75.9%	90.1%	11.0%	77.0%	82.3%	94.0%	4.8%	84.0%
M8	86.5%	89.0%	21.8%	82.5%	72.7%	85.4%	15.7%	72.4%	79.0%	88.5%	20.3%	76.6%
M9	90.6%	93.2%	18.4%	90.0%	76.1%	89.8%	17.4%	72.4%	81.9%	93.3%	8.7%	81.0%
S1	90.9%	93.4%	25.3%	88.8%	76.4%	89.4%	18.0%	75.7%	83.5%	94.9%	20.7%	72.0%
S2	84.0%	86.6%	5.7%	92.5%	73.0%	84.3%	12.8%	83.6%	74.2%	82.5%	6.6%	88.9%
S3	90.9%	93.1%	32.2%	91.3%	76.6%	88.8%	23.8%	74.3%	83.9%	95.4%	22.4%	70.5%

Column **Mod ID** is the same of Table 1. Columns *Acc*, *Pos*, *Neut*, and *Neg* represent the accuracy computed on all the reviews, on only the positive ones, on only the neutral ones, and only the negative ones respectively, for the three datasets of interest **Castles**, **Restaurants**, and **Hotels**. The table is best interpreted row by row to identify the most effective models. Models with overall better performance are indicated by a higher concentration of green cells, while those with poorer performance have more red cells. For example, notice that models M6 and M8 have a higher quantity of red across their rows, suggesting they perform worse compared to other models

Precision. The frequency with which they misclassified neutral or negative reviews as positive ones seems negligible compared to the number of correctly classified positive reviews. Concerning neutral polarity, the results from Table 3 are corroborated. As indicated by the *Recall*, the models struggle to identify neutral reviews, and furthermore, they frequently misclassified positive and negative reviews as neutral. This challenge may arise from the inherent difficulty in detecting true neutral sentiment, as human users seldom express sentiments that are entirely neutral. Even reviews that may appear neutral to a human reader often contain a mix of positive and negative aspects regarding the location of interest, making it challenging for LLMs to interpret them as truly neutral.

Finally, the findings regarding negative polarity are particularly noteworthy. Despite the high *Recall*, suggesting proficiency in recognising negative reviews, the limited *Precision* implies that the models often misclassified positive and neutral reviews as negative. This phenomenon is especially pronounced in the **Castles** dataset, previously noted for its anomaly wherein models found it easier to recognise negative reviews than positive ones. This suggests a tendency among models to classify positive reviews originating from this dataset as negative. However, it is important to note that the low *Precision* could also be partially attributed to the higher imbalance between negative and positive reviews in the **Castles** dataset, potentially leading to a higher number of false negatives due to the greater number of positive reviews. It is noteworthy to notice that most of the SOTA models appears to present slightly higher *Precision* than the zero-shot counterparts when considering the negative polarity in all the dataset of interest.

Based on the presented results, we opted to deepen our experimental evaluation by excluding neutral reviews from the datasets of interest, focusing solely on recognising positive and negative polarities. As we would expect, Table 5 reveals an observable increase in average accuracy, attributable to the simplified task of

Table 4 Results of the sentiment analysis task when considering reviews with a score of three stars as neutral

Mod ID	Castles			Restaurants			Hotels			Sent
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	
M1	0.990	0.925	0.956	0.931	0.867	0.898	0.950	0.948	0.949	*Positive*
M2	0.993	0.897	0.942	0.946	0.812	0.874	0.952	0.930	0.941	
M3	0.980	0.916	0.947	0.880	0.903	0.891	0.900	0.923	0.911	
M4	0.987	0.926	0.955	0.899	0.892	0.895	0.921	0.933	0.927	
M5	0.983	0.939	0.961	0.856	0.902	0.878	0.888	0.940	0.913	
M6	0.986	0.841	0.908	0.891	0.845	0.868	0.908	0.869	0.888	
M7	0.980	0.934	0.956	0.879	0.901	0.890	0.895	0.940	0.917	
M8	0.985	0.890	0.935	0.867	0.854	0.860	0.901	0.885	0.893	
M9	0.978	0.932	0.955	0.867	0.898	0.882	0.894	0.933	0.913	
S1	0.982	0.934	0.957	0.882	0.894	0.888	0.901	0.949	0.924	
S2	0.984	0.866	0.921	0.913	0.843	0.877	0.925	0.825	0.872	
S3	0.985	0.931	0.957	0.903	0.888	0.895	0.901	0.954	0.927	
M1	0.125	0.195	0.152	0.297	0.128	0.179	0.412	0.098	0.159	*Neutral*
M2	0.135	0.345	0.194	0.315	0.273	0.293	0.397	0.166	0.234	
M3	0.300	0.103	0.154	0.388	0.110	0.172	0.423	0.050	0.090	
M4	0.208	0.057	0.090	0.529	0.052	0.095	0.650	0.033	0.062	
M5	0.176	0.069	0.099	0.300	0.070	0.113	0.390	0.073	0.123	
M6	0.119	0.138	0.128	0.262	0.099	0.143	0.323	0.103	0.157	
M7	0.235	0.092	0.132	0.380	0.110	0.171	0.341	0.048	0.084	
M8	0.128	0.218	0.162	0.325	0.157	0.212	0.276	0.203	0.234	
M9	0.203	0.184	0.193	0.366	0.174	0.236	0.330	0.087	0.138	
S1	0.152	0.253	0.190	0.352	0.180	0.238	0.333	0.207	0.255	
S2	0.035	0.057	0.043	0.328	0.128	0.184	0.169	0.066	0.095	
S3	0.175	0.322	0.227	0.342	0.238	0.281	0.364	0.224	0.278	
M1	0.425	0.988	0.594	0.452	0.901	0.602	0.528	0.963	0.682	*Negative*
M2	0.438	0.963	0.602	0.456	0.855	0.595	0.536	0.954	0.687	
M3	0.242	0.888	0.381	0.469	0.783	0.586	0.486	0.854	0.619	
M4	0.263	0.950	0.412	0.415	0.855	0.559	0.488	0.937	0.642	
M5	0.313	0.938	0.469	0.450	0.717	0.553	0.532	0.777	0.631	
M6	0.175	0.913	0.294	0.388	0.763	0.514	0.409	0.823	0.546	
M7	0.297	0.913	0.448	0.459	0.770	0.575	0.536	0.840	0.655	
M8	0.265	0.825	0.401	0.437	0.724	0.545	0.545	0.766	0.637	
M9	0.360	0.900	0.514	0.514	0.724	0.601	0.548	0.810	0.654	
S1	0.507	0.888	0.645	0.511	0.757	0.610	0.695	0.720	0.707	
S2	0.239	0.925	0.379	0.401	0.836	0.542	0.372	0.889	0.525	
S3	0.525	0.913	0.667	0.521	0.743	0.612	0.703	0.705	0.704	

Columns *P*, *R* and *F1* represent the *Precision*, *Recall* and *F Score* measurements respectively. Column Sent represent the sentiment considered to compute the measurements. For the other Columns refer to Table 3

identifying only two possible polarities. Moreover, these results generally align with our conjectures regarding the models under scrutiny: although on the **Castles** and **Hotels** datasets, models *M1* and *M2* appear to outperform others. Furthermore, as we would expect due to its fine-tuning focused on distinguishing between the positive and negative polarities, *S4* outperforms all the other models when we consider the overall accuracy. However, it is interesting to notice that when the accuracy is computed only over the negative reviews, *M1* and *M2* still present slightly better performances.

Table 6 provides further validation of the findings presented in Table 4, indicating that the models demonstrate more reliable recognition of positive sentiment.

Table 5 Results of the sentiment analysis task when removing reviews with a score of three stars from the datasets

Mod ID	Castles			Restaurants			Hotels		
	Acc	*Pos*	*Neg*	*Acc*	*Pos*	*Neg*	*Acc*	*Pos*	*Neg*
M1	96.1%	96.0%	98.8%	89.9%	89.0%	94.1%	96.1%	95.9%	97.4%
M2	95.6%	95.6%	97.5%	88.9%	87.8%	94.7%	95.5%	95.3%	97.1%
M3	91.9%	92.0%	90.0%	90.4%	91.9%	82.9%	92.0%	92.9%	86.0%
M4	93.1%	93.0%	96.3%	88.8%	89.4%	85.5%	93.5%	93.5%	93.8%
M5	94.6%	94.7%	93.8%	89.5%	92.5%	74.3%	93.0%	95.0%	79.4%
M6	86.5%	86.2%	95.0%	86.4%	86.3%	86.8%	88.3%	88.7%	85.6%
M7	93.8%	93.9%	92.5%	90.6%	92.1%	82.9%	93.7%	95.0%	84.7%
M8	92.9%	93.0%	90.0%	87.7%	89.3%	79.6%	92.8%	94.4%	82.0%
M9	95.0%	95.1%	91.3%	90.7%	93.0%	78.9%	93.6%	94.8%	84.7%
S1	93.2%	93.4%	88.8%	87.2%	89.4%	75.7%	92.0%	94.9%	72.0%
S2	86.8%	86.6%	92.5%	84.2%	84.3%	83.6%	83.3%	82.5%	88.9%
S3	93.0%	93.1%	91.3%	86.4%	88.8%	74.3%	92.2%	95.4%	70.5%
S4	97.3%	97.3%	96.3%	93.0%	93.0%	92.8%	97.7%	98.2%	94.1%

The Columns are the same presented in Table 3. The table should be read row by row to identify the best models. Green cells indicate better performance, while red cells indicate poorer performance. Models like M1 and M2 show higher concentrations of green cells, suggesting overall better performance, whereas models like M6 show lower performance, highlighted by a higher presence of red cells.

The observed increase in *Precision*, particularly evident in the **Restaurants** and **Hotels** datasets compared to Table 4, suggests that a considerable number of neutral reviews may have been previously misclassified as negative. The relatively modest increase in *Precision* for the **Castles** dataset could further corroborate the hypothesis that the greater imbalance between positive and negative reviews may, to some extent, contribute to the lower *Precision* observed. *S4* still best performances overall even if, interestingly, *S1* and *S3* present significantly higher *Precision* when applied to the **Castles** dataset.

Overall, our experimental evaluation indicates that, particularly when focusing on recognising positive and negative polarity, the models exhibit sufficient reliability for the task at hand. However, the pronounced imbalance between the number of positive and negative reviews characteristic of such datasets renders it challenging to accurately evaluate the *Precision* and, consequently, the *F Score* of the models for both positive and negative polarities. Concerning the performance differences between the zero-shot and SOTA models, the latter appears to perform slightly better than the former. However, when evaluating the entire dataset, this difference does not seem to be substantial, with the exception of model *S*4. The specialisation of *S*4 in distinguishing positive and negative polarity allows it to excel when neutral reviews are excluded.

5.2 Keywords extraction

In this sub-section we present our results considering the keyword extraction task. In Figs. 6, 7, 8, 9, 10, and 11, we depict the medians and IQ ranges computed across the responses of our models of interest for the entire datasets. These histograms

Table 6 Results of the sentiment analysis task when removing reviews with a score of three stars from the datasets

Mod ID	Castles			Restaurants			Hotels			Sent
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	
M1	1.000	0.960	0.979	0.987	0.890	0.936	0.996	0.959	0.977	*Positive*
M2	0.999	0.956	0.977	0.988	0.878	0.930	0.996	0.953	0.974	
M3	0.996	0.920	0.957	0.965	0.919	0.941	0.978	0.929	0.953	
M4	0.999	0.930	0.963	0.969	0.894	0.930	0.990	0.935	0.962	
M5	0.998	0.947	0.972	0.948	0.925	0.937	0.969	0.950	0.959	
M6	0.998	0.862	0.925	0.971	0.863	0.914	0.977	0.887	0.930	
M7	0.997	0.939	0.967	0.965	0.921	0.943	0.977	0.950	0.963	
M8	0.996	0.930	0.962	0.957	0.893	0.924	0.973	0.944	0.958	
M9	0.997	0.951	0.974	0.958	0.930	0.944	0.977	0.948	0.963	
S1	0.997	0.934	0.964	0.967	0.894	0.929	0.983	0.949	0.966	
S2	0.998	0.866	0.927	0.981	0.843	0.906	0.989	0.825	0.899	
S3	0.999	0.931	0.964	0.980	0.888	0.932	0.984	0.954	0.969	
S4	0.999	0.973	0.986	0.985	0.930	0.957	0.991	0.982	0.987	
M1	0.451	0.988	0.620	0.627	0.941	0.753	0.776	0.974	0.864	*Negative*
M2	0.424	0.975	0.591	0.603	0.947	0.737	0.751	0.971	0.847	
M3	0.274	0.900	0.420	0.667	0.829	0.739	0.639	0.860	0.733	
M4	0.317	0.963	0.477	0.613	0.855	0.714	0.678	0.938	0.787	
M5	0.371	0.938	0.532	0.661	0.743	0.700	0.697	0.794	0.742	
M6	0.188	0.950	0.313	0.555	0.868	0.677	0.525	0.856	0.651	
M7	0.336	0.925	0.493	0.674	0.829	0.743	0.711	0.847	0.773	
M8	0.300	0.900	0.450	0.593	0.796	0.680	0.679	0.820	0.743	
M9	0.386	0.913	0.543	0.690	0.789	0.736	0.706	0.847	0.770	
S1	0.657	0.888	0.755	0.752	0.757	0.754	0.901	0.720	0.800	
S2	0.288	0.925	0.439	0.588	0.836	0.690	0.496	0.889	0.637	
S3	0.664	0.913	0.768	0.774	0.743	0.758	0.888	0.705	0.786	
S4	0.546	0.963	0.697	0.723	0.928	0.813	0.884	0.941	0.912	

The Columns are the same presented in Table 4.

illustrate the distribution of samples presenting median and IQ ranges within certain value intervals for each specific keyword. The objective is to observe whether the models exhibit unanimity and confidence in their selection of the keyword, as indicated by polarised medians towards 0 (indicating a mismatch) or 1 (indicating a match). Conversely, the spread of the likelihood predicted by the models, as reflected by the IQ range, should primarily tend towards zero, as higher IQ ranges suggest strong disagreement among the models regarding the likelihood. As can be seen, the results are strictly dependant on the particular keyword and dataset considered, and they do not seem to always indicate high confidence in the prediction provided by the models considered. The most significant example of this behaviour seems to be "tour" for the **Castle** Dataset which seems to present an almost uniform distribution for the medians and a nearly Gaussian distribution centred between 0.4 and 0.5 for the IQ ranges.

For most of the other keywords and datasets, the results align with our expectations, with some noteworthy exceptions. Particularly, we observe that the keyword "view" in the **Castle** Dataset, chosen with the assistance of the related word cloud, exhibits a distribution of IQ ranges that is less skewed than desired. However, it is intriguing to note that, compared to the "panorama" keyword, we observe a significantly higher number of samples with very high medians. This indicates that while the "panorama" keyword was infrequently matched to reviews with a high degree of

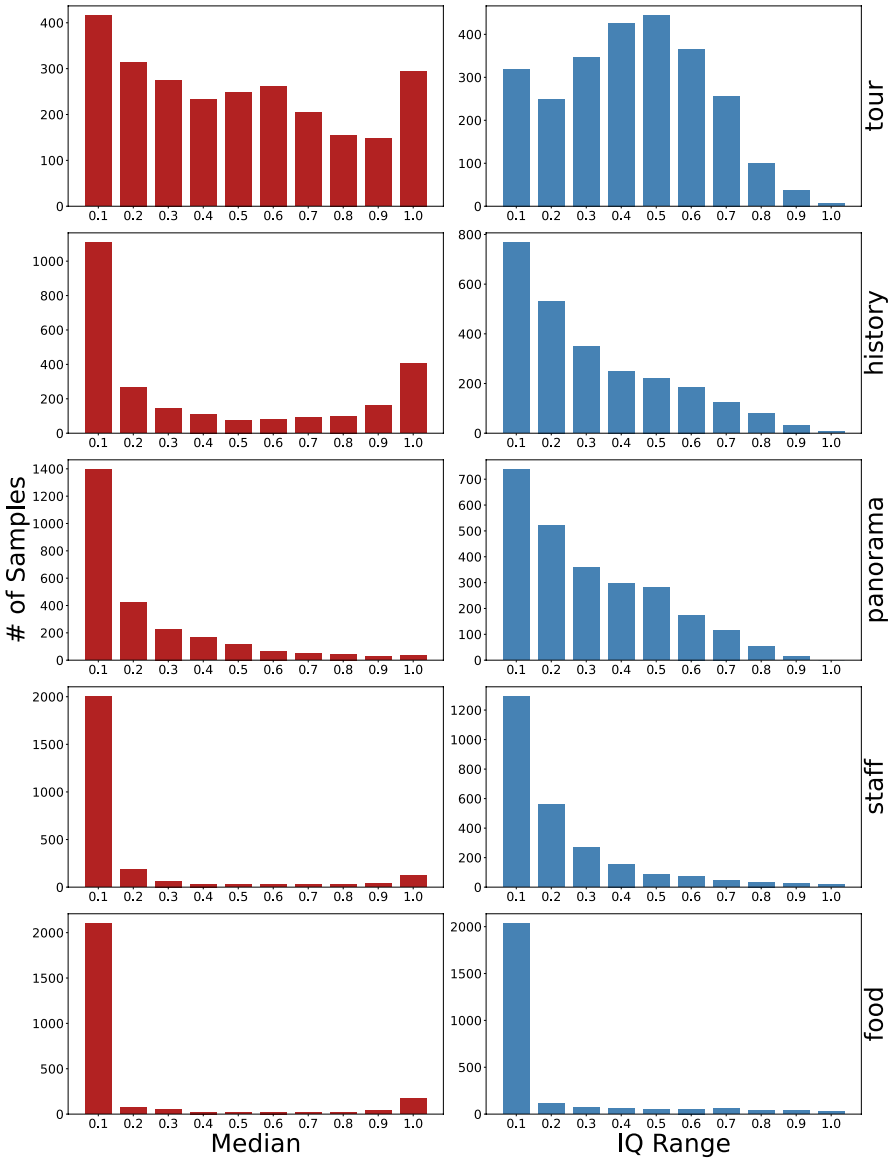

Fig. 6 Histograms illustrating the sample distribution and the median and interquartile range of probabilities calculated by the models for the **Castles** Dataset across the different keywords of SET0

certainty, the "view" keyword is matched to more samples but with a lower degree of certainty. It is important to emphasise that while we seek high certainty in our predictions, a high degree of certainty regarding a keyword that rarely appears in the reviews of interest may not provide much useful information. Contrarily, a keyword

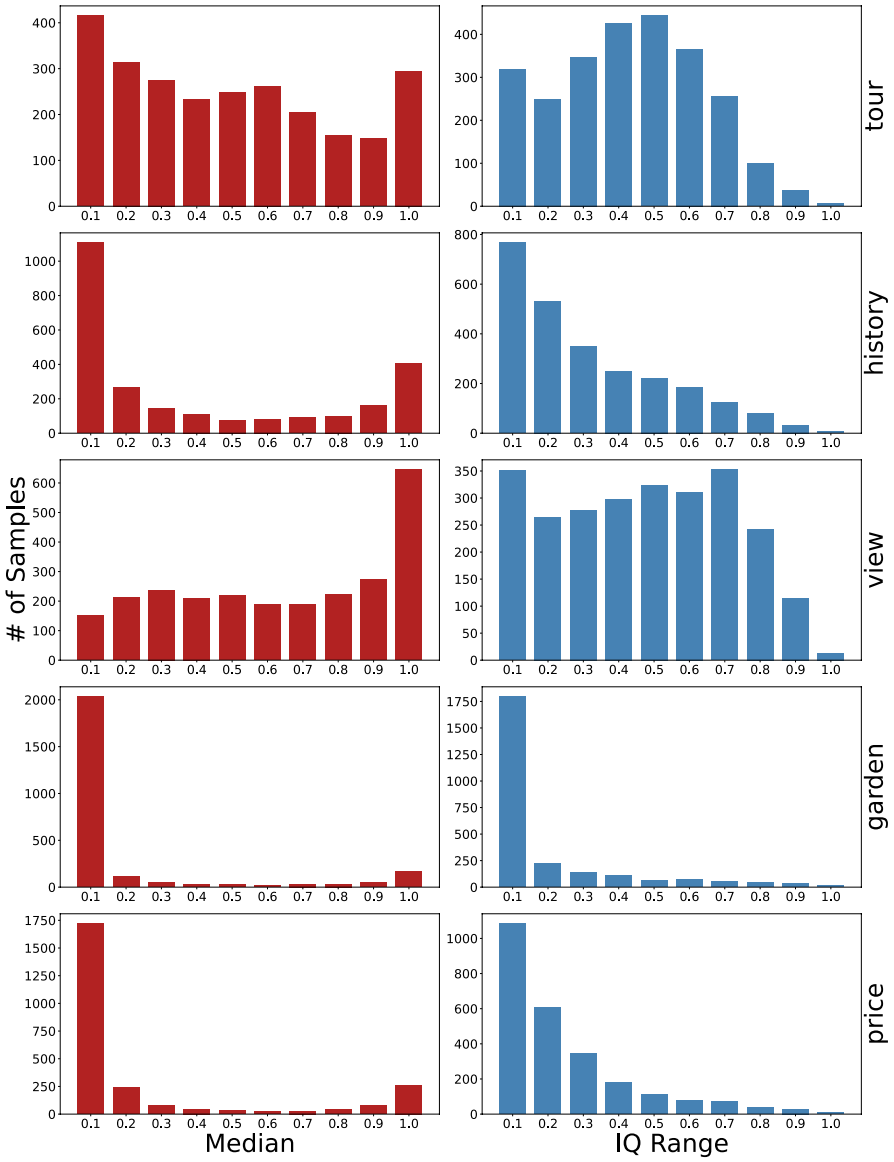

Fig. 7 Histograms illustrating the sample distribution and the median and interquartile range of probabilities calculated by the models for the **Castles** Dataset across the different keywords of SET1

that is identified less reliably but more frequently may still offer valuable insights to the end user.

Regarding the difference in results obtained for the keywords of *SET1* and those of *SET2*, it does not appear that the models of interest consistently provide more reliable or relevant results for keywords selected with the assistance of the word

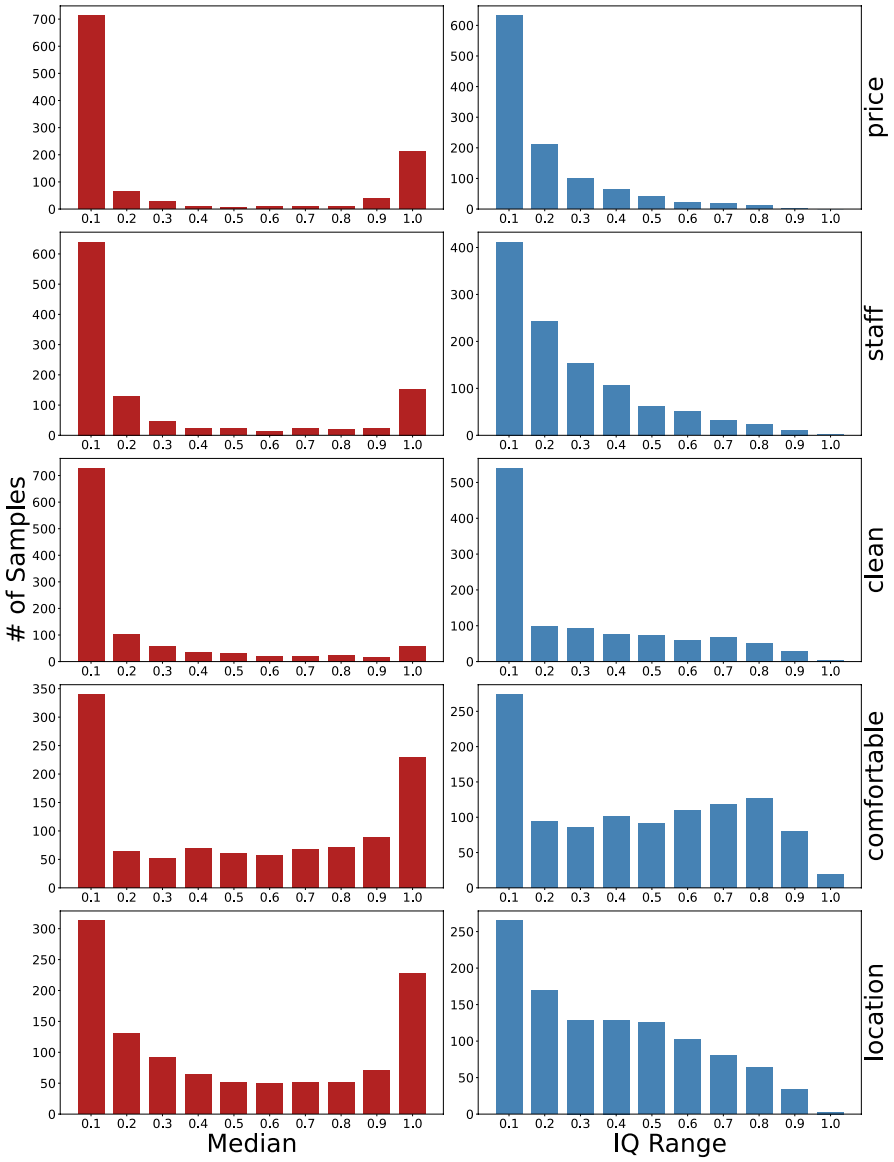

Fig. 8 Histograms illustrating the sample distribution and the median and interquartile range of probabilities calculated by the models for the **Restaurants** Dataset across the different keywords of SET0

clouds. However, specific cases demonstrate instances where the keywords chosen with the support of the word cloud yield significantly better results than the original ones. Two such examples are the keywords “breakfast” and “clean” in the **Hotels** Dataset. As depicted in Fig. 10 and 11, both of these keywords exhibit distributions of medians that are more polarised towards 0 and 1, as well as distributions of IQ

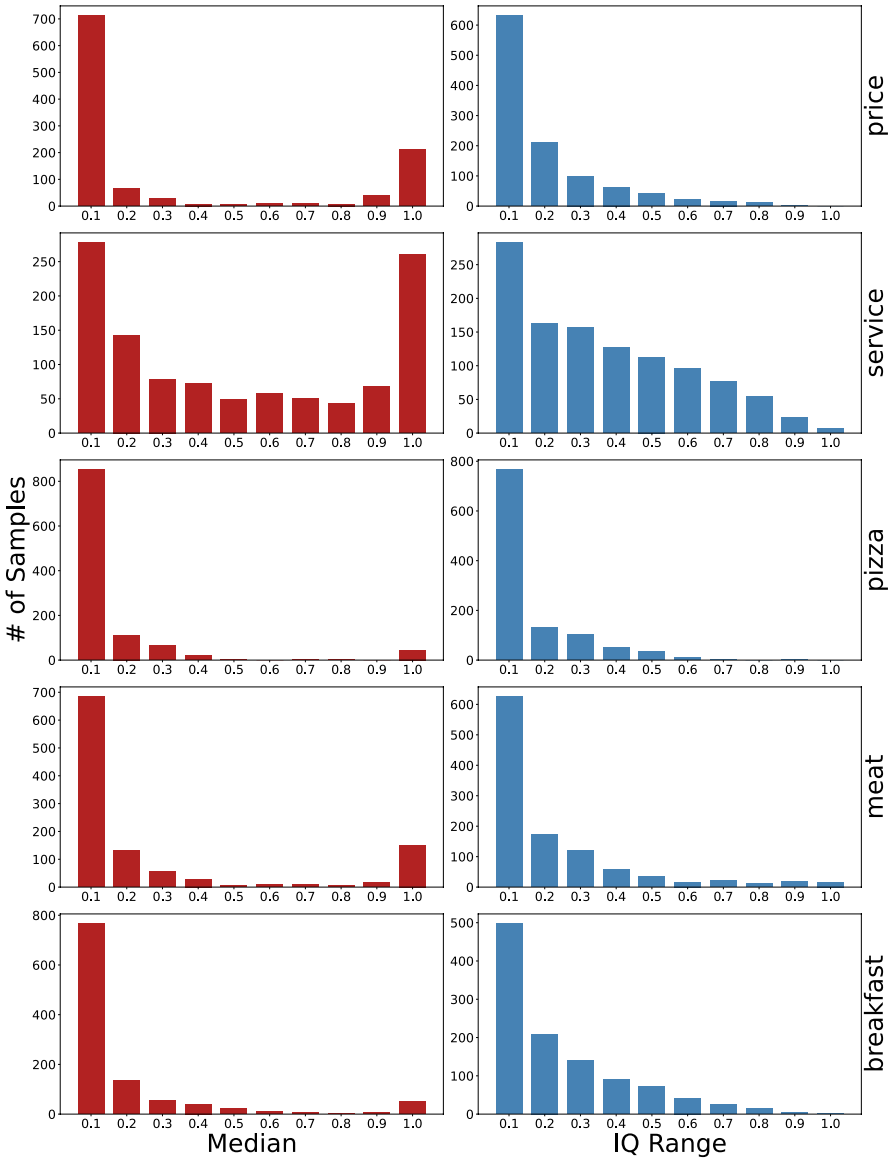

Fig. 9 Histograms illustrating the sample distribution and the median and interquartile range of probabilities calculated by the models for the **Restaurants** Dataset across the different keywords of SET1

ranges that are more skewed towards 0 compared to the original keywords "food" and "ambience".

We conclude our experimental evaluation with some considerations regarding the behaviour of the models of interest when applied to specific reviews.

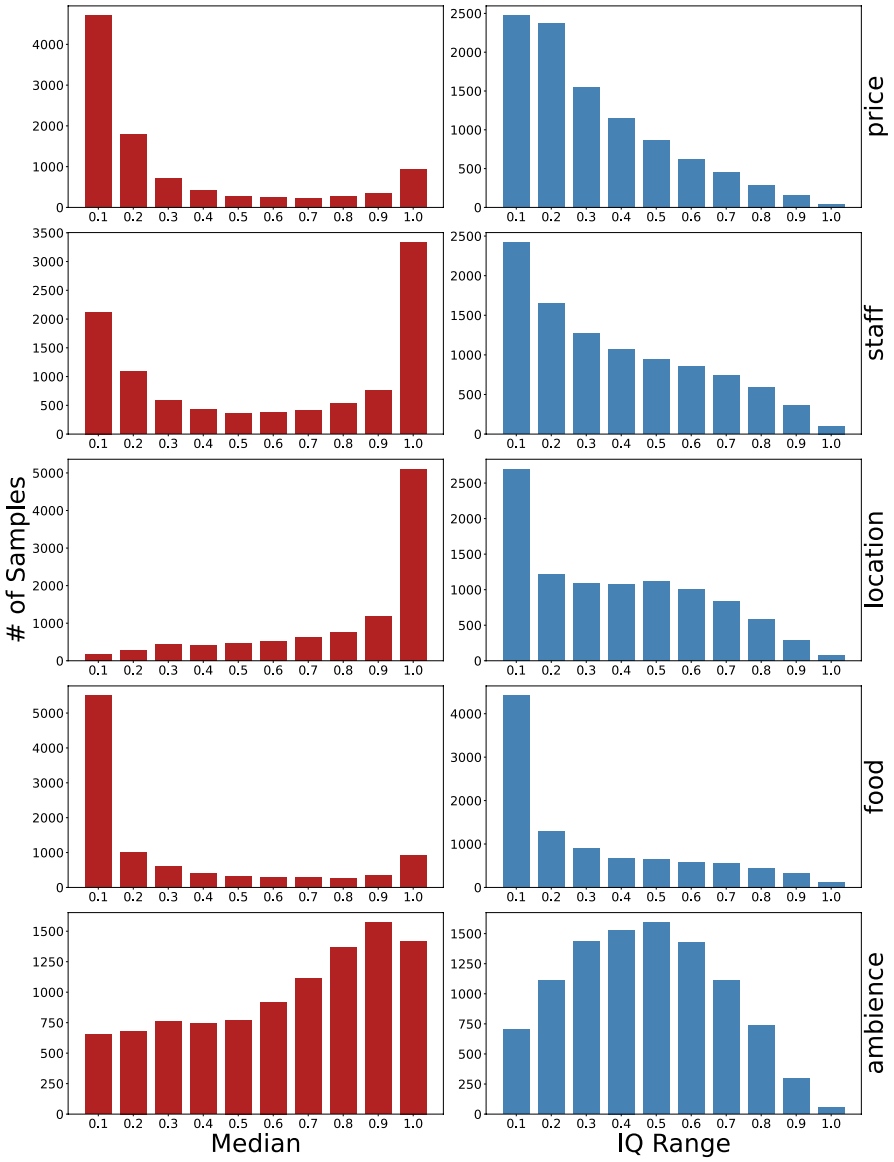

Fig. 10 Histograms illustrating the sample distribution and the median and interquartile range of probabilities calculated by the models for the **Hotels** Dataset across the different keywords of SET0

Specifically, we focus briefly on the first and third reviews of the **Castle** Dataset, which we believe can provide valuable insights. As depicted in Fig. 12, the models exhibit varying degrees of agreement regarding which keywords should be

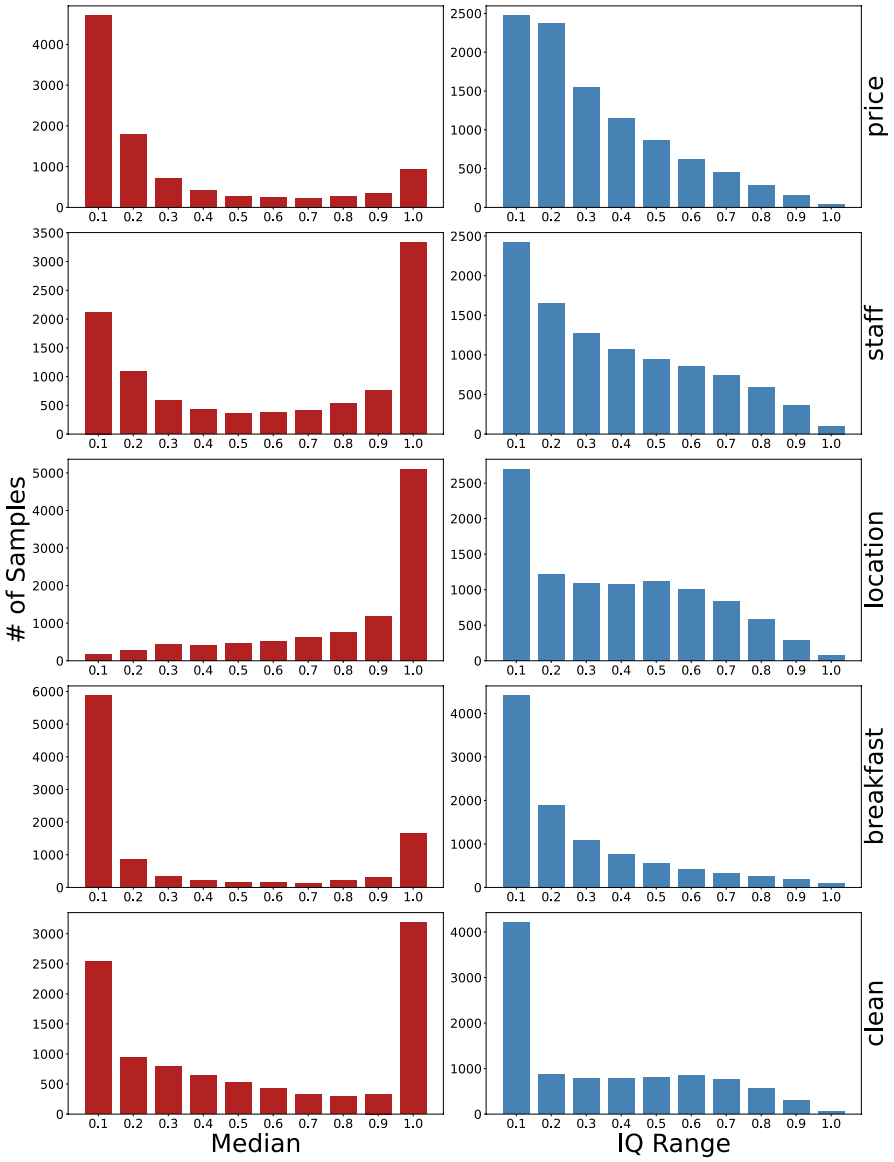

Fig. 11 Histograms illustrating the sample distribution and the median and interquartile range of probabilities calculated by the models for the **Hotels** Dataset across the different keywords of SET1

matched to the first review, whereas they appear to be more confident in pairing the third review with the keywords "staff" and "food". To understand the rationale behind these results, we need to examine the reviews in question more closely:

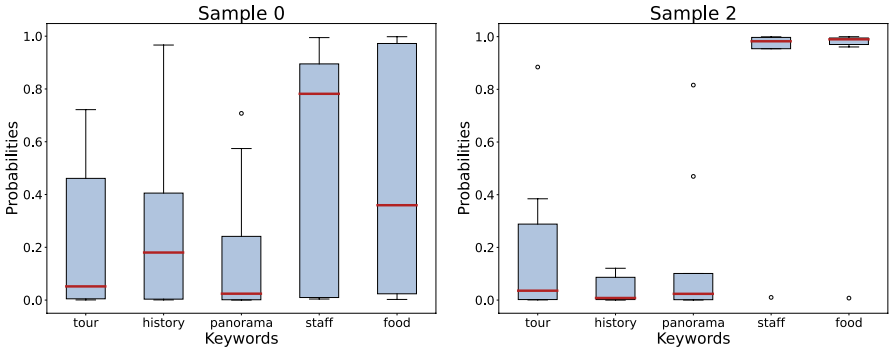

Fig. 12 Boxplots illustrating the probabilities for each keyword in *SET1* predicted by the models of interest when applied to the first (Sample 0) and third (Sample 2) samples of the **Castle** Dataset, respectively

- Sample 0: "*More than a great and magical place. Clean,* **tasty** *, unique for walks, suitable for children ... there was a big party downstairs in the* **restaurants** *... everything was again very clean,* **tasty**, *wonderful ... the* **staff is great polite and smiling**, *... The experience with the ducks and the swan is very nice!*".
- Sample 2: "*Very* **tasty large cup of cappuccino** *and* **very good fresh ice cream**. *And for a 5 star hotel in a castle on an island on a river for very reasonable and affordable price. Oh yes, and* **very friendly staff**.".

Firstly, it is important to note that we have provided only brief excerpts of Sample 0 for the sake of conciseness. The complete review spans 192 words and describes various aspects of the user's stay at the hotel associated with the castle. Conversely, Sample 2 is presented in its entirety as it comprises only 39 words.

While the behaviour of the models under consideration is generally not formally explainable, we can posit some hypotheses by comparing the two reviews. Specifically, we observe that the word "tasty", which we would semantically associate with the keyword "food", appears to be used as a generic adjective to denote a positive sentiment in the first review. This unexpected usage of the term, coupled with the presence of the term "restaurants", may contribute to the models' uncertainty regarding the keyword "food". Even to human observers, it is not entirely clear whether the "tasty" mentioned in the first review refers to some type of food consumed by the reviewer at the castle or if it is simply an unintended use of the term, possibly attributable to English being a second language for the reviewer or to errors in machine translation applied to the review in its original language. Conversely, in the second review, the reference to food is much clearer: the user mentions a "tasty large cup of cappuccino" and "very good fresh ice cream", leading to a more unanimous and certain match with the keyword "food" by the models.

The variance in the models' confidence levels regarding the keyword "staff" appears to arise from different factors. Both reviews explicitly mention the staff of the respective locations: the first review states "... the staff is great, polite, and smiling...", while the second mentions "... very friendly staff". Consequently, it is

plausible that the disparity in the models' confidence levels may be attributed to the length discrepancy between the two reviews – the first being nearly five times longer than the second – and the discussion of various events at the hotel, present in the first review, which the models struggle to interpret. Specifically, the reviewer recounts instances of noise disturbances occurring at different times of the day, a narrative that necessitates a certain level of contextual understanding, a challenge for LLMs, which are known for their difficulty in replicating common sense reasoning. As a result, this narrative segment of the review appears to complicate the models' comprehension of the text, consequently leading to lower confidence levels in their predictions.

Overall, our experimental assessment revealed that while NLP technologies can aid domain experts in selecting sets of relevant keywords, their usage should be supplemented with human oversight to prevent the selection of sub optimal keywords. Additionally, we verified that the models under scrutiny can reliably match reviews with keywords of interest to a certain extent. However, the overall reliability of these models may fluctuate depending on the specific keywords involved. Lastly, our analysis of two particular reviews highlighted that the confidence of the models' predictions appears to be significantly influenced by both the length and writing style of the reviews.

6 Conclusions

In this paper, we focused on showcasing how LLMs' capabilities can enhance decision-making and service delivery in the tourism industry. Particularly, we explored two applications we believe could be of interest for tourism enterprises: sentiment analysis and keyword extraction. We examined models that are readily available and do not necessitate additional training, which could be prohibitively costly in terms of resources and thus unfeasible for SMEs. In the subsequent sub-sections, we will discuss the theoretical contribution of the study, analyse the practical implications of the presented research in the tourism domain, examine the limitations of LLMs identified through our research, and outline potential future research directions in this domain.

6.1 Theoretical contribution

This paper contributes theoretically by exploring the capabilities of LLMs in sentiment analysis and keyword extraction using zero-shot learning within the domain of tourism and hospitality. Our experimental evaluation encompassed three distinct datasets and thirteen models, demonstrating their ability to identify positive and negative sentiments in the majority of reviews analysed. However, the significant imbalance between positive and negative reviews in the datasets posed challenges in precisely evaluating the models' *Precision* in sentiment analysis. Overall, the models showed effectiveness in aligning reviews with keywords identified by domain

experts, though their success varied depending on the specific keywords. Notably, keywords supported by NLP techniques often yielded better alignment with the reviews. Further analysis of individual reviews provided insights into how factors such as review length, writing style, and language nuances impact the models' predictive accuracy. In summary, this paper highlights how integrating LLM technologies can empower tourism enterprises, including SMEs, to extract actionable insights from customer feedback more efficiently, thereby enhancing their competitive edge in the market. This research can contribute to expanding the understanding of LLM applications in tourism and lays a foundation for future studies leveraging advanced NLP techniques for analysing tourism-related data.

6.2 Practical implications

From a practical perspective, the findings of our research can offer relevant insights for SMEs in the tourism domain. By using LLMs for sentiment analysis and keyword extraction, tourism managers can gain a deeper understanding of customer sentiments and preferences with minimal computational resources. Additionally, the use of open-source LLM architectures democratises access to advanced analytical tools, enabling even resource-constrained organisations to benefit from cutting-edge technology. Our collaboration with SMEs within a technology transfer project has demonstrated the practical feasibility of integrating LLMs into everyday business practices. The positive feedback received from these collaborations underlines the value of tailored solutions in meeting specific business needs, thereby enhancing operational efficiency and customer engagement. Such partnerships not only validate the applicability of LLM-driven insights but also foster ongoing innovation and adaptation in response to industry demands. This mutual exchange of knowledge and expertise ensures that SMEs can harness the full potential of LLM technologies to drive sustainable growth and competitive advantage in the tourism sector.

6.3 Limitations of LLMs

Despite the promising results, it is essential to acknowledge the limitations of LLMs in this context. One primary limitation is the variability in model performance across different datasets and keywords, indicating that LLMs may not consistently provide accurate results in all scenarios. Additionally, the lack of a ground truth for keyword extraction poses challenges in evaluating the models' performance rigorously. Furthermore, LLMs can sometimes struggle with neutral sentiment detection and may exhibit biases inherent in the training data used to develop them. These constraints emphasise the importance of interpreting the results carefully and pinpoint areas where future advancements in LLM development can be focused.

6.4 Future directions for research

Firstly, our immediate objective involves creating meticulously annotated datasets specifically designed for training purposes. These datasets will not only aid in refining keyword extraction models but also strengthen the reliability of model assessments, even in scenarios where ground truth data is unavailable. Secondly, there is a need to develop more robust evaluation metrics for keyword extraction tasks that do not rely on ground truth data. Enhancing these metrics would significantly improve the accuracy and trustworthiness of model evaluations. Furthermore, exploring the performance of LLMs across a broader range of datasets and within diverse tourism contexts is essential. This broader exploration will provide a more comprehensive understanding of how LLMs can be effectively applied in real-world scenarios within the tourism industry. Additionally, there is a critical need for research aimed at mitigating biases inherent in LLMs and enhancing their capability to accurately detect neutral sentiments. Addressing these challenges will be instrumental in advancing the reliability and fairness of LLM applications across various domains. These future directions are crucial for overcoming current limitations and advancing the field of NLP within the tourism sector and beyond.

Acknowledgements This work has been developed within the framework of the project e.INS- Ecosystem of Innovation for Next Generation Sardinia (cod. ECS 00000038) funded by the Italian Ministry for Research and Education (MUR) under the National Recovery and Resilience Plan (NRRP) - MISSION 4 COMPONENT 2, "From research to business" INVESTMENT 1.5, "Creation and strengthening of Ecosystems of innovation" and construction of "Territorial R&D Leaders", CUP J83C21000320007.

Author contributions All authors whose names appear on the submission: 1. made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; 2. drafted the work or revised it critically for important intellectual content; 3. approved the version to be published; 4. agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Data availability and access The datasets used in this work are freely available on Kaggle and can be found at the following links: 1. European Castles Dataset: https://www.kaggle.com/datasets/datasciencedonut/european-castles (License: CC0 (https://creativecommons.org/publicdomain/zero/1.0/)). 2. Google Maps Restaurant Reviews Dataset: https://www.kaggle.com/datasets/denizbilginn/google-maps-restaurant-reviews (License: ODbL (https://opendatacommons.org/licenses/odbl/1-0/)). 3. Hotel Reviews Dataset: https://www.kaggle.com/datasets/datafiniti/hotel-reviews (License: CC BY-NC-SA 4.0(https://creativecommons.org/licenses/by-nc-sa/4.0/)).

Code availability The code required to replicate our experiments is available online at https://github.com/AIMet-Lab/ITT-2024-PNRR-ZSSA.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical and informed consent for data used No informed consent is needed for the data used.

References

Abdullah T, Ahmet A (2023) Deep learning in sentiment analysis: recent architectures. ACM Comput Surv 55(8):159:1-159:37. https://doi.org/10.1145/3548772

Ameur A, Hamdi S, Yahia SB (2024) Sentiment analysis for hotel reviews: a systematic literature review. ACM Comput Surv 56(2):51:1-51:38. https://doi.org/10.1145/3605152

Augenstein I, Das M, Riedel S, et al (2017) Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017. Association for Computational Linguistics, pp 546–555, https://doi.org/10.18653/V1/S17-2091

Bagherzadeh S, Shokouhyar S, Jahani H et al (2021) A generalizable sentiment analysis method for creating a hotel dictionary: using big data on tripadvisor hotel reviews. J Hosp Tour Technol 12(2):210–238. https://doi.org/10.1108/JHTT-02-2020-0034

Barbieri F, Anke LE, Camacho-Collados J (2021) XLM-T: a multilingual language model toolkit for twitter. CoRR abs/2104.12250

Barbieri F, Camacho-Collados J, Anke LE, et al (2020) Tweeteval: unified benchmark and comparative evaluation for tweet classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, Findings of ACL, vol EMNLP 2020. Association for Computational Linguistics, pp 1644–1650, https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.148

Birjali M, Kasri M, Hssane AB (2021) A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl Based Syst 226:107134. https://doi.org/10.1016/J.KNOSYS.2021.107134

Bowman SR, Angeli G, Potts C, et al (2015) A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. The Association for Computational Linguistics, pp 632–642, https://doi.org/10.18653/V1/D15-1075

Bucur C (2015) Using opinion mining techniques in tourism. Proc Econ Finance 23:1666–1673. https://doi.org/10.1016/S2212-5671(15)00471-2

Dagan I, Glickman O, Magnini B (2005) The PASCAL recognising textual entailment challenge. In: Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers, Lecture Notes in Computer Science, vol 3944. Springer, pp 177–190, https://doi.org/10.1007/11736790_9

de Souza JGR, de Paiva Oliveira A, de Andrade GC, et al (2018) A deep learning approach for sentiment analysis applied to hotel's reviews. In: Natural language processing and information systems-23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings, Lecture Notes in Computer Science, vol 10859. Springer, pp 48–56, https://doi.org/10.1007/978-3-319-91947-8_5

Devlin J, Chang M, Lee K, et al (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019,

Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp 4171–4186, https://doi.org/10.18653/V1/N19-1423

Farisi AA, Sibaroni Y, Faraby SA (2019) Sentiment analysis on hotel reviews using multinomial naÃ¯ve bayes classifier. J Phys: Conf Ser 1192(1):012024. https://doi.org/10.1088/1742-6596/1192/1/012024

Firoozeh N, Nazarenko A, Alizon F et al (2020) Keyword extraction: Issues and methods. Nat Lang Eng 26(3):259–291. https://doi.org/10.1017/S1351324919000457

Ghorpade T, Ragha L (2012) Featured based sentiment classification for hotel reviews using nlp and Bayesian classification. In: 2012 International Conference on Communication, Information & Computing Technology (ICCICT), pp 1–5, https://doi.org/10.1109/ICCICT.2012.6398136

Gräbner D, Zanker M, Fliedl G, et al (2012) Classification of customer reviews based on sentiment analysis. In: Information and Communication Technologies in Tourism 2012, ENTER 2012, Proceedings of the International Conference in Helsingborg, Sweden, January 25-27, 2012. Springer, pp 460–470, https://doi.org/10.1007/978-3-7091-1142-0_40

Hartmann J, Heitmann M, Siebert C et al (2023) More than a feeling: accuracy and application of sentiment analysis. Int J Res Mark 40(1):75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

He P, Liu X, Gao J, et al (2021) DeBERTa: decoding-enhanced bert with disentangled attention. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net

Hnin CC, Naw N, Win A (2018) Aspect level opinion mining for hotel reviews in Myanmar language. In: IEEE International Conference on Agents, ICA 2018, Singapore, July 28-31, 2018. IEEE, pp 132–135, https://doi.org/10.1109/AGENTS.2018.8460040

Jardim SVB, Mora C (2021) Customer reviews sentiment-based analysis and clustering for market oriented tourism services and products development or positioning. In: CENTERIS 2021 - International Conference on ENTERprise Information Systems / ProjMAN 2021 - International Conference on Project MANagement / HCist 2021 - International Conference on Health and Social Care Information Systems and Technologies 2021, Braga, Portugal, Procedia Computer Science, vol 196. Elsevier, pp 199–206, https://doi.org/10.1016/J.PROCS.2021.12.006

Jiang M, Zhang W, Zhang M et al (2019) An LSTM-CNN attention approach for aspect-level sentiment classification. J Comput Methods Sci Eng 19(4):859–868. https://doi.org/10.3233/JCM-190022

Kim SN, Medelyan O, Kan M, et al (2010) Semeval-2010 task 5 : automatic keyphrase extraction from scientific articles. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010. The Association for Computer Linguistics, pp 21–26

Laurer M, Van Atteveldt W, Casas A et al (2023) Less annotating, more classifying: addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. Polit Anal. https://doi.org/10.1017/pan.2023.20

Lewis M, Liu Y, Goyal N, et al (2020) BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, pp 7871–7880, https://doi.org/10.18653/V1/2020.ACL-MAIN.703

Liu Y, Ott M, Goyal N, et al (2019) RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692

Loureiro D, Barbieri F, Neves L, et al (2022) Timelms: diachronic language models from twitter. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, pp 251–260, https://doi.org/10.18653/V1/2022.ACL-DEMO.25

Martins GS, de Paiva Oliveira A, Moreira A (2017) Sentiment analysis applied to hotels evaluation. In: Computational Science and Its Applications - ICCSA 2017 - 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part VI, Lecture Notes in Computer Science, vol 10409. Springer, pp 710–716, https://doi.org/10.1007/978-3-319-62407-5_52

Nguyen DQ, Vu T, Nguyen AT (2020) Bertweet: A pre-trained language model for English tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020. Association for Computational Linguistics, pp 9–14, https://doi.org/10.18653/V1/2020.EMNLP-DEMOS.2

Pal S, Ghosh S, Nag A (2018) Sentiment analysis in the light of LSTM recurrent neural networks. Int J Synth Emot 9(1):33–39. https://doi.org/10.4018/IJSE.2018010103

Pencarelli T (2020) The digital revolution in the travel and tourism industry. J Inf Technol Tour 22(3):455–476. https://doi.org/10.1007/S40558-019-00160-3

Priyantina RA, Sarno R (2019) Sentiment analysis of hotel reviews using latent Dirichlet allocation, semantic similarity and lstm. Int J Intell Eng Syst 12(4):142–155. https://doi.org/10.22266/ijies2019.0831.14

Raiaan MAK, Mukta MSH, Fatema K et al (2024) A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access 12:26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

Rosenthal S, Farra N, Nakov P (2017) Semeval-2017 task 4: sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017. Association for Computational Linguistics, pp 502–518, https://doi.org/10.18653/V1/S17-2088

Rossetti M, Stella F, Zanker M (2016) Analyzing user reviews in tourism with topic models. J Inf Technol Tour 16(1):5–21. https://doi.org/10.1007/S40558-015-0035-Y

Sharma S, Diwakar M, Joshi K, et al (2022) A critical review on sentiment analysis techniques. In: 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), IEEE, pp 741–746, https://doi.org/10.1109/ICIEM54221.2022.9853140

Shi H, Li X (2011) A sentiment analysis model for hotel reviews based on supervised learning. In: International Conference on machine learning and cybernetics, ICMLC 2011, Guilin, China, July 10-13, 2011, Proceedings. IEEE, pp 950–954, https://doi.org/10.1109/ICMLC.2011.6016866

Tan KL, Lee C, Anbananthen KSM et al (2022) RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access 10:21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828

Tesfagergish SG, Kapočiūtė-Dzikienė J, Damaševičius R (2022) Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. Appl Sci 12(17):8662. https://doi.org/10.3390/app12178662

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp 5998–6008

Wang X, Ye Y, Gupta A (2018) Zero-shot recognition via semantic embeddings and knowledge graphs. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society, pp 6857–6866, https://doi.org/10.1109/CVPR.2018.00717

Wang W, Zheng VW, Yu H et al (2019) A survey of zero-shot learning: settings, methods, and applications. ACM Trans Intell Syst Technol 10(2):13:1-13:37. https://doi.org/10.1145/3293318

Wang W, Bao H, Huang S, et al (2021) Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In: Findings of the association for computational linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, Findings of ACL, vol ACL/IJCNLP 2021. Association for Computational Linguistics, pp 2140–2151, https://doi.org/10.18653/V1/2021.FINDINGS-ACL.188

Wang H, Li J, Wu H et al (2022) Pre-trained language models and their applications. Engineering 25:51–65. https://doi.org/10.1016/j.eng.2022.04.024

Wankhade M, Rao ACS, Kulkarni C (2022) A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55(7):5731–5780. https://doi.org/10.1007/S10462-022-10144-1

Williams A, Nangia N, Bowman SR (2018) A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). Association for Computational Linguistics, pp 1112–1122, https://doi.org/10.18653/V1/N18-1101

Xian Y, Lampert CH, Schiele B et al (2019) Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans Pattern Anal Mach Intell 41(9):2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768

Younas A, Nasim R, Ali S, et al (2020) Sentiment analysis of code-mixed roman Urdu-English social media text using deep learning approaches. In: 23rd IEEE International Conference on Computational Science and Engineering, CSE 2020, Guangzhou, China, December 29, 2020 - January 1, 2021. IEEE, pp 66–71, https://doi.org/10.1109/CSE50738.2020.00017