Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

Arabic Short-Text Dataset for Sentiment Analysis of Tourism and Leisure Events

Seham Basabain[1] | Ahmed Al-Dubai[2] | Erik Cambria[3] | Khalid Alomar[1] | Amir Hussain[2]

[1]Faculty of Computing and Information Technology, King AbdulAziz University, Jeddah, Saudi Arabia | [2]School of Computing, Edinburgh Napier University, Edinburgh, UK | [3]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

Correspondence: Seham Basabain (ssbasabain@kau.edu.sa; seham.basabain@napier.ac.uk)

Received: 10 March 2024 | **Revised:** 1 September 2024 | **Accepted:** 10 February 2025

Funding: Amir Hussain would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC)—Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1.

Keywords: Arabic sentiment analysis | automatic labelling | Saudi tourism | twitter | zero-shot learning

ABSTRACT

The focus of this study is to present the detailed process of collecting a dataset of Arabic short-text in the tourism context and annotating this dataset for the task of sentiment analysis using an automatic zero-shot labelling technique utilising transformer-based models. This is benchmarked against a baseline manual annotation approach utilising native Arab human annotators. This study also introduces an approach exploiting both manual/handcrafted and automatically generated annotations of the dataset tweets for the task of sentiment analysis as part of a cross-domain approach using a model trained on sarcasm labels and vice versa. The total collected corpus size is 2293 tweets; after annotation, these tweets were labelled in a three-way classification approach as either positive, negative or neutral. We run different experiments to provide benchmark results of Arabic sentiment classification. Comparative results on our dataset show that the highest performing baseline model when utilising manual labels was MARBERT, with an accuracy of up to 87%, which was pre-trained for Arabic on a massive amount of data. It should be noted that this model enhanced its performance additionally after pre-training on a dialectical Arabic and modern standard Arabic corpus. On the other hand, zero-shot automatically generated labels achieved an 84% accuracy rate in predicting sarcasm classes from sentiment labels.

1 | Introduction

With the growth of social media platforms, such as Twitter, there has been a significant increase in user-generated content related to various topics, including travel and tourism. These platforms allow users to share their experiences, opinions, and recommendations about different tourist destinations. Social media analytics and artificial intelligence (AI) have proven to be effective tools in extracting valuable information from these vast amounts of user-generated content (Liu 2020). However, most of the existing studies have focused on collecting and analysing data in English or other widely spoken languages. This poses a challenging problem when it comes to analysing and understanding the sentiments and preferences of Arabic-speaking users, who correspond to 4.8% of all Internet users, which makes Arabic the fourth most used language after English, Chinese and Spanish (Oueslati et al. 2020). The lack of a large number of Arabic datasets with valuable information hinders the development and application of sentiment analysis techniques for Arabic social media content (Oueslati et al. 2020).

In an attempt to address this gap, this study aims to collect a dataset from Twitter specifically in the Arabic language, focusing on tweets related to tourism. The dataset will be gathered by selecting relevant keywords and hashtags related to tourism in Arabic and collecting a sufficient number of tweets from various

users. The dataset will be manually labelled by a team of human annotators who are fluent in Arabic and have knowledge about the tourism domain. In addition to the manual labelling approach, an automatic labelling technique will also be employed to compare and evaluate its effectiveness in capturing sentiment information.

The process of labelling data is crucial in many research areas, particularly in the field of natural language processing (NLP). Traditionally, manual labelling has been the predominant method used to label datasets. However, manual labelling can be time-consuming and resource-intensive, as it requires human annotators to go through each data point and assign class labels based on their understanding of the content. This process can be subjective and prone to human biases.

To address these challenges, researchers have explored alternative approaches such as zero-shot automatic labelling (Chakravarthy and Arutla 2022). Zero-shot automatic labelling is a technique that aims to automatically assign class labels to data instances without relying on any pre-defined sentiment labels or training data specific to a certain domain, which is particularly valuable when working with large datasets. Instead, this technique saves time and resources that would have been required for manual labelling by leveraging the power of transfer learning to generate labels using pre-trained language models.

Language models, such as BERT (Devlin et al. 2019), have been trained on large amounts of text data, allowing them to understand and generate human-like language. These language models can be fine-tuned to perform zero-shot classification, where they are trained on related tasks using general language data.

The main contribution of this study is to fill the gap in Arabic sentiment analysis by creating and utilising a dataset of 2293 Arabic tweets in the tourism domain. This dataset is crucial for understanding the preferences and sentiments of Arabic-speaking users. Unlike existing studies that predominantly rely on manually labelled data, our approach uniquely combines manual and zero-shot automatic labelling techniques. This dual approach enhances the efficiency and scalability of sentiment analysis, reduces human biases and provides a robust framework for analysing Arabic social media content. We annotated the dataset with three labels—positive, negative and neutral—using both methods and compared their performance with different baseline models. This innovation sets our study apart from prior research as it pioneers the use of zero-shot learning for Arabic sentiment analysis, demonstrating its effectiveness without domain-specific training data. The corpus is publicly available for the research community.[1]

The paper is structured as follows. Section 2 gives a thorough representation of previous work related to Arabic dataset generation and gives an overview of some of the benchmark datasets used in Arabic literature. Section 3 describes the process of generating, filtering, and manually annotating our collected dataset. Section 4 presents the zero-shot automatic labelling approach and how to utilise language models as zero-shot labellers. Section 5 compares the performance of both approaches in terms of sentiment classification accuracy. Section 6 discusses the final results then the final section concludes the work and suggests some points for future work.

2 | Related Work

Sentiment analysis is a classification process of linguistics that analyses perceptions, feelings and emotions. Literature on sentiment classification focuses on online opinion resources, including review sites, blogs, and social media posts, accessible news outlets and discussion forums (Ramanathan and Meyyappan 2019). Unlike most data resources, Twitter messages are very short, up to 140 characters, and include various languages, local dialects, informal slang, and spelling errors. Therefore, some sentiment analysis methods and feature selection approaches may not work properly for Twitter, like other sources, and vice versa (Aldakhil 2020).

Although it is a widely studied research area, Arabic sentiment analysis is still considered to be a low-resource language, unlike English, where many benchmark datasets and approaches exist for the task (Guellil et al. 2019). In sentiment analysis, a text is classified as positive, negative, or neutral; according to Kirilenko et al. (2018), many obstacles might arise when tackling this task, and the development of classification algorithms is still an uncharted research area. In general, current Arabic sentiment classification research employs different approaches ranging from simple machine learning dictionary-based algorithms to complex deep learning approaches.

A work by Alomari et al. (2017) presented the Arabic Jordanian General Tweets (AJGT) a publicly available corpus containing 1800 annotated tweets in a two-way classification approach as either positive or negative. These collected tweets are written in Jordanian dialects and modern standard Arabic from numerous general themes. In their work, they examined machine learning supervised methods for the task of Arabic sentiment analysis by comparing two machine learning algorithms: the Support Vector Machine and Naïve Bayes, with numerous attributes and preprocessing approaches.

Rosenthal et al. (2017) introduced SemEval 2017, where the tweets in this dataset are classified in a three-way classification approach into positive, negative, or neutral. The dataset contains three separate sets: a training set of 3355, a validation set of 671, and a test set of 6100 Arabic tweets. All tweets were collected through specifying some topics notable during the period of collection, and these topics are different in training and test sets since each set was collected in a different period. All tweets were annotated using the Appen crowdsourcing platform, previously known as CrowdFlower.

Another SemEval task for affect classification in Arabic tweets was introduced in 2018 (Mohammad et al. 2018). In this version, five subtasks are publicly available of novel social media texts, where systems can be utilised to automatically determine the affective state of a person from his tweet. These five subtasks are: Emotion Intensity Regression (EI-reg), Emotion Intensity Ordinal Classification (EI-oc), Valence (Sentiment) Regression (V-reg), Valence Ordinal Classification (V-oc) and Emotion

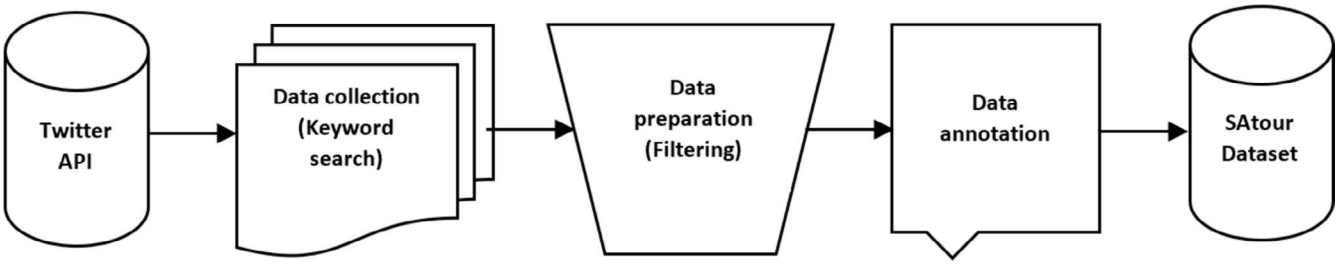

FIGURE 1 | Dataset creation methodology.

classification (E-c). For each subtask, there are three separate training, validation, and testing labelled sets of different languages: Arabic, English, and Spanish.

A work by Nabil et al. (2015) has presented the Arabic Sentiment Tweets Datasets (ASTD), an Arabic social sentiment analysis dataset collected from Twitter. This corpus comprises almost 10 k tweets labelled as objective, subjective positive, subjective negative, or subjective mixed. The study provides a comprehensive presentation of the dataset's characteristics and statistical information. Additionally, it includes experimental findings for four-way sentiment classification, obtained utilising the dataset's predefined partitions.

Another large corpus for Arabic sentiment analysis is the Arabic Speech Act and Sentiment corpus of tweets (ArSAS) (Elmadany et al. 2018). It is the largest Arabic speech behaviour and emotion annotation corpus covering multiple topics of 21 k labelled tweets with six speech act labels and four sentiment labels. Furthermore, it is believed to be the first Arabic letter tagged to identify a speech act in tweets. According to Al-Twairesh et al. (2016), the corpus is expected to increase awareness of the Arabic speech behaviour recognition project and stimulate more research into Arabic sentiment analysis. The presence of multiple dialects in this dataset challenges the classification problem. In simpler datasets with only one dialect, such as ASTD, supervised learning algorithms will be able to identify the top keywords associated with each emotion in that dialect. When there are multiple dialects, the range of keywords associated with each mood increases because the words vary from one dialect to another.

In the field of tourism, sentiment analysis can also be utilised for determining tourists' emotions towards the services and places they visit by investigating the big data available online instead of using traditional sampling and statistical analysis methods. It has been observed that social media, tourism blogs and websites are the core attributes of data that scholars use to elicit tourists' perceptions or opinions about specific places they visit (Ainin et al. 2020). Alkhaldi et al. (2022) explored the possibility of investigating the contribution of different sentiment analysis approaches and compared their performance in the field of travel text classification. Nevertheless, the application of sentiment analysis in the travel industry faces considerable challenges (Alharbi et al. 2022). Numerous sentiment classification methods have been proposed for English, Japanese and Chinese. Therefore, research on automatic sentiment analysis must be performed in other languages for different natural language applications (Flores-Ruiz et al. 2021).

3 | Dataset Creation

In this section, the methodology of constructing the dataset is explained. Figure 1 illustrates this adopted methodology consisting of three phases: data collection, data preparation, and data annotation. These phases are further described below.

3.1 | Data Collection

Tweets on a list of topics were collected and filtered using the Twitter API.[2] Zhao and Jiang (2011) defined three topical categories as follows:

- Long-Standing: topics that have been frequently discussed for a considerable amount of time.
- Entity: Topics pertaining to people or groups.
- Event: A significant event that is taking place.

These types of topics were chosen in relation to the tourism domain in Saudi Arabia, including long-standing events such as the seasons occurring in main cities, entities like those tweets about certain singers or celebrities visiting the kingdom, and events currently occurring in different interests. Between late 2021 and early 2022, a set of random tweets was added to the corpus belonging to specific keywords related to Saudi seasons of tourism in different cities of Saudi Arabia, including Riyadh, Jeddah, Jizan, and AlUla. These keywords include popular events of Saudi historical, entertaining, sport, musical events, and spots. Table 1 lists the keywords used for scraping tweets.

Another criterion applied to collect tweets was the language of the text in tweets, where only Arabic tweets were collected using the 'ar' label in the API to scrape tweets authored in Arabic.

3.2 | Data Preparation and Filtering

From the first phase, we managed to collect a set of 23,500 tweets; we applied some criteria to filter these data and prepare them for the next phase. First, we removed all English characters, including URLs, to analyse results on Arabic text solely. Second, all duplicate tweets were removed, including retweets. Additionally, twitter special characters were removed, like hashtags '#' and account mentions '@'. Tweets containing advertisements, political opinions and supplications were excluded. Moreover, tweets including media were excluded since most of these tweets were found to be spam.

TABLE 1 | Keywords for scraping Tweets.

# موسم_الرياض	# ونتر_لاند	# واجهة_الرياض	# فعاليات_موسم_الرياض	# بوليفارد_رياض_سيتي
# آرت_بروميناد	# سيتي_ووك	# موسم_جدة	# تخيل_أكثر	# هيئة_الترفيه
# ساوندستورم	# ميدل_بست	# يوم_بدينا	# يوم_التنكر	# أيامنا_الحلوة
# العلا	# موضي_الشمراني	# شتاء_جازان	# ايقاف_موسم_الرياض_مطلب	# فورمولا1_في_السعودية
			# مهرجان_جازان_شتوي	# يوم_التأسيس

TABLE 2 | SAtour tweets examples.

Tweet	English translation	Label
موسم الرياض موسم عالمي ونجاح سعودي	Riyadh season is a global season and a Saudi success	Positive
تخيلو اضيّع وقتي هناك، للامانه مايروح الا المراهقين ميدل بست	Imagine wasting my time there, to be honest, only teenagers go to the Middle Beast	Negative
السلام عليكم ... هل يتطلب شراء تذكره لطفل أصغر من سنتين لحضور مهرجان الرياض للألعاب الذي سيقام في واجهة الرياض؟	Peace be upon you... Is it required to purchase a ticket for a child under two years of age to attend the Riyadh Games Festival, which will be held at the Riyadh Front?	Neutral

TABLE 3 | The fleiss kappa coefficient κ of SAtour manual annotation.

No. annotators per tweet	No. of tweets			κ
$\eta = 3$	Agree upon all three annotators	Agree upon two annotators	Not agreed upon all annotators	0.7986
	1857 (80.99%)	389 (16.96%)	47 (2.05%)	

Inclusion criteria were to keep tweets of Arabic text only. These tweets should be about a topic of the Saudi tourism domain only, enquiries about events or holding an opinion about events. All Arabic dialects are included, not only Saudi, because some of these tweets were written by non-Saudi visitors; although the majority of the extracted text was found to be either in MSA or gulf dialect. The final dataset, after applying the above cleaning and filtering steps, had 2293 tweets.

3.3 | Dataset Manual Annotation

The objective of this phase is to have all tweets annotated based on tweet-level into a three-way classification of sentiments, where a tweet can be either positive, negative or neutral. The annotation process was undertaken by three Saudi natives. Annotation guidelines and the challenges during the annotation process were highlighted to provide insights for annotators. Following the work of Al-Twairesh et al. (2016) the annotation guidelines were as follows:

- All annotations should be considered from the author's perspective, not the annotator.
- News or enquiries should be annotated as neutral since they convey no sentiment.
- If a tweet contains mixed positive and negative content, this tweet should be excluded.

These annotations were chosen based on majority voting following the work of Al-Twairesh et al. (2016). An example of some of these annotated tweets with their translation is provided in Table 2.

With a ratio of three annotations per tweet, we calculated the Inter-Annotator Agreement (IAA) to measure the reliability of the labels. For that, we adopted the Fleiss Kappa coefficient κ (Fleiss 1971) which can be calculated as:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}; where \tag{1}$$

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{2}$$

$$\overline{P}_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{n} n_{ij} \right)^2 \tag{3}$$

where k is the number of classes, N is the total number of tweets, n is the number of annotations per each tweet and n_{ij} is the number of annotators who agreed on a tweet i to be assigned to a certain class j. In our case, we had $N = 3$, $n = 3$, the overall κ The Fleiss Kappa coefficient of SAtour is 0.8, which indicates that the agreement among the annotators was substantial. Table 3 shows the agreement level upon all annotators assigned to the task.

The results of the Inter-annotator agreement reveal that 81% of the total number of tweets were agreed upon by all three annotators with the same label, 17% had an agreed label from two annotators, while the last 2% of the tweets were not agreed upon

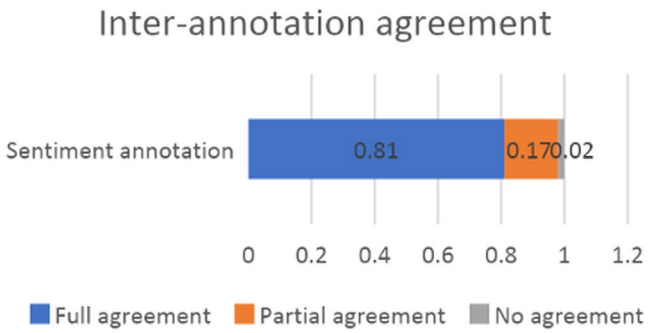

FIGURE 2 | Percentage of Tweets with full, partial, or no agreement on sentiment labels.

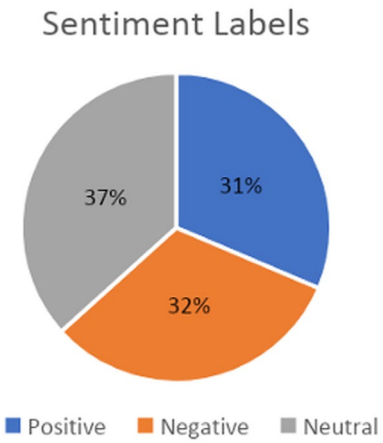

FIGURE 3 | SAtour sentiment labels statistics.

by all annotators. Figure 2 provides a visualisation of these percentages.

Final dataset has 720 (31%) Positive, 731 (32%) Negative, and 842 (37%) Neutral tweets, as shown in Figure 3.

Moreover, a dialect identification analysis was performed on SAtour utilising (*CAMeL-Lab/bert-base-arabic-camelbert-msa-did-madar-twitter5*).[3] Figure 4 illustrates all identified dialects in SAtour with their coverage ratio. Analysis results reveal that the Saudi dialect was dominant among collected text with up to 1845 tweets and formulates 80.46% of the dataset. The Kuwaiti dialect comes in second place with 7% dataset coverage and a collection of 162 tweets.

4 | Zero-Shot Automatic Labelling for SAtour

Zero-shot learning has been tackled as an interesting topic for recent research. This is due to its ability to scale classification models without the need for further training data. This is typically accomplished by connecting classes with semantic information such as characteristics (Annadani and Biswas 2018). The idea was originally proposed by (Radford et al. 2019) and it is similar to few-shot learning; the only difference is to provide the model with a description of the task in a natural language way instead of any labelled examples of inputs (Brown et al. 2020).

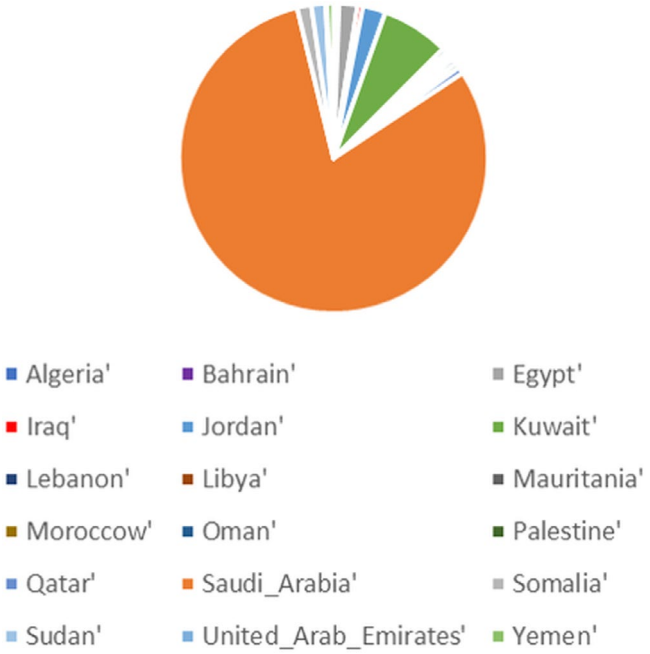

FIGURE 4 | SAtour dialect identification analysis.

The goal of zero-shot learning is to detect items whose instances were not seen during training, in which the label space for training sets and test sets is disconnected from one another and there are no samples to train test classes. This problem of having a disjoint test and train label sets in a zero-shot learning environment can be addressed by tackling related sub-problems, such as learning intermediate attribute classifiers and learning a mixture of seen class proportions, or by a direct approach, such as compatibility learning frameworks.

Furthermore, for difficult classification tasks, zero-shot learning is proposed to overcome the issues that hampered the advancement of traditional systems. The first issue is about annotating large-scale samples, which is both time-consuming and expensive; the second issue is that new categories are constantly emerging and some of them are difficult or even dangerous to collect, such as the identified coffin fish in the deep sea. In contrast, zero-shot learning has the advantage of addressing data annotation and class recognition problems, making it an increasingly common subject lately (Xie et al. 2019).

Additionally, there are certain tagged training examples in the feature space for zero-shot learning. The classes associated with these instruction examples are known as the seen classes. There are a few unlabelled testing situations in the feature space that correspond to a different assortment of classes known as the unseen classes. The feature space is typically a real number space, and each instance is displayed as a vector within it and is often related to a single class (Wang, Zheng, et al. 2019). This semantic space is a high-dimensional vector space that links both seen and unseen classes and allows for the knowledge transfer of seen classes to unseen classes.

Zero-shot learning originally emerged in the field of computer vision, where models are taught to detect objects or categories that do not exist in the training data (Han et al. 2022). In order to use this strategy, a mapping between object visual attributes

and their semantic representations must be learned. This concept was built upon and applied to text-based tasks in NLP (Pourpanah et al. 2022).

4.1 | Automated Data Labelling

Automatic dataset labelling is the process of automatically assigning applicable and precise labels to significant volumes of data (Mehrotra et al. 2013). This unique method speeds up the labour-intensive and time-consuming process of labelling datasets using machine learning algorithms and artificial intelligence approaches. Businesses can drastically save the time and labour needed to annotate data by automating this process (Khalel and El-Saban 2018).

Multiple research investigations on how to automatically label data have separated the job into two major areas: first is the candidate label generation and second is the candidate label rating. The candidate label generation area is concerned with creating words to represent the labels for the generated topic model clusters, while the candidate ranking concentrates on ranking the generated labels based on a certain scoring approach (Khan and Chua 2021).

As an instance, Kapadia (2019) creates candidate labels by requesting the Wikipedia API with ranking topic parameters to obtain the headings and sub-headings of articles written on Wikipedia. The rankings are then computed via a model with supervised learning based on association metrics and linguistic traits. Because their model is deliberately trained on blogs, news, books and PubMed, it may perform well at some tasks while underperforming elsewhere (Khan and Chua 2021).

The study conducted by (Alqarafi et al. 2018) focused on building a semi-supervised corpus annotation for Saudi sentiment analysis using Twitter data. The researchers recognised the potential of Twitter as a valuable source of data for sentiment analysis, particularly in the context of Saudi Arabia. In order to construct the corpus, the authors targeted a set of sentiment words and used them to extract tweets containing these words. The authors aimed to overcome the challenges of limited labelled data in Arabic sentiment analysis by adopting a semi-supervised approach. This approach allowed them to leverage both manual annotation and automated techniques to annotate a large corpus of Saudi tweets with sentiment labels. To evaluate the effectiveness of their approach, the authors compared their semi-supervised corpus annotation with existing sentiment analysis corpora.

Guellil et al. (2018) introduced Sentialg, which is an automated corpus annotation tool specifically designed for Algerian sentiment analysis. To accomplish this, the researchers utilised the BosonNLP toolkit and trained it using the Weibo corpus for annotation and training purposes.

The use of sophisticated machine learning models is the basis of automatic dataset labelling (Kastrati et al. 2020). For the purpose of finding patterns, correlations and semantic interpretations in the data, these models are trained using labelled datasets. Once trained, these models can be used to infer or attribute the correct labels to unlabelled data based on the learned patterns and insights (Almuqren and Cristea 2021).

Automatic dataset labelling has a number of advantages, such as increased efficiency, decreased costs and quicker model development, but it also presents some challenges. It is crucial to ensure the quality and accuracy of automatically assigned labels, as labelling errors can negatively affect subsequent tasks and models. Maintaining high-quality labels requires evaluation, continuous feedback cycles and human oversight (McCann et al. 2018).

4.2 | Language Models as Zero-Shot Learners for Automatic Labelling

In NLP, pre-trained language models, such as BERT (Devlin et al. 2019), may generalise well to different natural language understanding tasks and capture detailed semantic information. These language models compute the weight of the model using either intra-attention or global attention. The former takes any two words in a sentence and estimates the similarity between them. On the other hand, global attention focuses on the entire textual content (Fernández-Isabel et al. 2023). In fact, zero-shot automatic labelling techniques can utilise this by assigning labels to unlabelled instances based on their semantic similarity to a specified set of labels by utilising the learned representations (Rezaei and Shahidi 2020).

Zero-shot learning has shown promising results in various domains. Authors have used zero-shot learning in automatic dataset labelling in NLP. Zero-shot automatic labelling is a technique that aims to label unlabelled data using a classification model without any explicit training on the specific task or dataset (Wang, Tang, et al. 2019). Traditional automatic labelling methods require a significant amount of annotated data for training, which can be costly and time-consuming. Zero-shot automatic labelling, on the other hand, leverages the generalisation capability of pre-trained models to assign labels to new instances without the need for explicit training (Xu et al. 2020).

Zero-shot text classification techniques can be used for automated data labelling by modifying the task as a classification issue with K candidate labels as classes, with the candidate labels having the highest likelihood chosen as appropriate labels for the dataset of interest. This can be carried out for the whole dataset as well as for identified clusters (Åslund 2021).

Language models have emerged as a foundational technology for various NLP tasks, such as the automatic labelling of datasets (Bojanowski 2017). These models are intended to comprehend and generate text that resembles human language by learning the statistical patterns and semantic relationships inherent in immense quantities of textual data. They have revolutionised the field of NLP by providing potent tools for text classification, named entity recognition, sentiment analysis and other tasks (Misargopoulos et al. 2022).

Moreover, large pre-trained language models like GPT-3 have developed an unexpected ability to execute zero-shot learning.

To classify sentiment without any training samples, for example, we can prompt the language model with the review and label description 'Does the student like this subject?' and ask whether the following word is 'Yes' or 'No'. The next word prediction training objective, however, remains mismatched with the target zero-shot learning goal (Zhong et al. 2021). Moreover, the study of Zhao et al. (2023) applied different pre-trained models to several language understanding tasks without labelled or additional unlabelled data. As a result, pre-trained language models have proven effectiveness for a wide range of NLP tasks. Existing techniques, on the other hand, necessitate either fine-tuning on downstream labelled datasets or manually creating appropriate prompts. They offer nonparametric prompting pre-trained language models (NPPrompt) for fully zero-shot language understanding in their research. Unlike prior methods, NPPrompt solely uses pre-trained language models and does not require labelled data or extra raw corpus for fine-tuning, nor does it rely on humans to generate an extensive collection of label words (Zhao et al. 2023).

Automatic dataset categorisation benefits in multiple ways from language models. First, these models can be used for text classification, facilitating the labelling of unlabelled data based on learned representations and contextual understanding (Dube et al. 2019). This automated labelling process is especially beneficial for large datasets where manual labelling would be time consuming and expensive.

Additionally, language models excel at extracting pertinent information from text. They can recognise and assign named entities, sentiments, topics and other important characteristics within the dataset (Alsanad 2018). By leveraging pre-trained knowledge and fine-tuning on labelled data specific to the dataset labelling task, language models can effectively assign labels to instances that were not previously labelled.

Dube et al. (2019) present a technique for automatic dataset labelling to enable transfer learning. The authors demonstrate an approach where large unlabelled datasets are automatically labelled using deep learning models. They compute the distance between an unlabelled data point and the average response of these models to create 'pseudo labels' for the unlabelled data. These pseudo labelled datasets are then used to train source models for transfer learning. The authors evaluate different methods of pseudo labelling and compare the transfer learning accuracy achieved by models trained on these pseudo labelled datasets against models trained on human-annotated ImageNet1K labels.

The work of Liu et al. (2021) presents a promising approach for addressing the zero-shot multi-label text classification problem. Traditional methods utilise graph encoders to incorporate label hierarchies and learn matching models between text and labels. However, they have limited exploration of pre-trained models like BERT, which do not generate individual vector representations for text or labels. In this work, the authors propose a Reinforced Label Hierarchy Reasoning (RLHR) approach to enhance pre-trained models with label hierarchies during training. They also introduce a rollback algorithm to handle logical errors in predictions during inference. Additionally, Khan and Chua (2021) address the challenge of manually labelling topics generated by Latent Dirichlet Allocation (LDA) in a large corpus of documents. Manual labelling is time-consuming, expensive and subjective. To automate the labelling process, the authors propose a system framework that leverages pre-trained zero-shot classification models. They conduct experiments using news content as a dataset, focusing on five topic categories: Crime, Business, Entertainment, Science and Politics.

5 | Benchmark Experiments

During baseline experiments, the SAtour dataset was split into training, validation, and testing sets with the percentages of 80%, 10%, and 10% respectively; Table 4 presents the manual labels distribution in the data splits.

Three baseline models were utilised to test our dataset.

Traditional Baseline Models: In this approach, we applied TF-IDF + SVM to extract features from the input tweets; these features are extracted based on word importance representation with respect to the entire corpus. Then, an SVM model is utilised to classify these inputs.

Deep Learning Baseline Models: In this approach, we utilised BERT-based models pre-trained on Arabic text as feature extractors and classifiers. Recently, utilising pre-trained language word embeddings generated efficient feature vectors with less effort and prevented overfitting of the models (Mohammad et al. 2021). Thus, we fine-tuned AraBERT (*bert-base-arabertv02-twitter*) (Antoun et al. 2020) and MARBERT (Abdul-Mageed et al. 2021). Fine-tuning these pre-trained models requires a specific way to pre-process the data to have a fixed size of word IDs for each tweet. Moreover, these models were trained with a learning rate of 2×10^{-5} and all their default hyper parameters in the huggingface[4] library were transferred to our target task.

Zero/Few-Shot Models: In this approach, we have used (*arabic_xlm_xnli*)[5] as a classifier to automatically annotate collected tweets. This model is an *XLM-Roberta-base* model fine-tuned on the XNLI dataset in Arabic. It was basically developed for the task of zero-shot hate speech detection. We have utilised this pre-trained model with the same parameters defined in the huggingface library of a 2×10^{-5} learning rate, 32 batch size and a maximum length equal to 128. Moreover, we have applied (*sn-xlm-roberta-base-snli-mnli-anli-xnli*)[6] as a feature extractor. This model is also based on *XLM-Roberta-base*, but it was trained on four datasets including: SNLI, MNLI, ANLI and XNLI. The model is based on the Siamese network for

TABLE 4 | Manual sentiment labels distributions in data split.

Label	Train	Validation	Test	Total
Positive	568	75	77	720
Negative	586	66	79	731
Neutral	680	88	74	842
Total	1834	229	230	2293

zero/few-shot learning text classification. It uses sentence-transformers library to map sentences and paragraphs to a dimensional dense vector space. Then, the extracted features were fed to a prototypical network for sentiment prediction in a few-shot setting following the work of (Basabain et al. 2023). In this approach finding the mapping function between an input text and a class label is the primary objective of a prototypical network meta learner for text classification. Where training this meta learner requires an iteration over a series of episodic mini-batches.

Each mini-batch episode, labelled as $B_E = \{S, Q\}$, is created by selecting input text from D_{train} to form a support set $S = \{(s_1, y_1),...,(s_n, y_n)\}$ and a query set $Q = \{(q_1, y_1),...,(q_m, y_m)\}$ along with their corresponding true labels. The episodes are defined by three variables: the number of classes k, the number of instances per class in the support set $n = |S_k|$, and the number of examples per class in the query set $m = |Q_k|$. S_k and Q_k indicate the sets of text samples labelled as $k = y$ in the support set and the query set, respectively. The variables $\{k, n, m\}$ are hyperparameters that regulate the batch creation process and require precise adjustment. After generating the support and query sets, the comprehensive function f_θ is trained to extract contextual embedding vectors from these sets. f_θ establishes the encoding model with parameters θ. Subsequently, the mean of the embedding vectors from the support set corresponding to each class k is computed to provide the prototype vector c_k as:

$$c_k = \left(\frac{1}{|S_k|}\right) \cdot \sum_{(s_i, y_k) \epsilon S_k} f_\theta(s_i) \tag{4}$$

where the function f represents the (*sn-xlm-roberta-base-snli-mnli-anli-xnli*) feature extractor that transforms the input text into an embedding vector. In our problem, let (s_i, y_k) be an element that belongs to the support set S_k of class k. After creating prototypes for each class in the feature space, sentiment classification is carried out using a nearest prototype centroid approach as implemented by Prototypical networks, which use Euclidean distance to determine the d distance between each class prototype and the embedded query point as illustrated in this equation:

$$P_\theta\left(y = k \mid q_i\right) = \frac{\exp\left(-d\left(f_\theta(q_i), c_k\right)\right)}{\sum_{k'} \exp\left(-d\left(f_\theta(q_i), c_{k'}\right)\right)} \tag{5}$$

This is achieved by creating a probability distribution over all classes $k' = 1...|k|$ through a distance-based non-parametric classifier and a softmax function applied to the distances between all class prototype vectors and the target query vector q_i. The query vector is labelled based on the class of the closest prototype.

Another experiment in this approach implemented was to apply a cross-domain few-shot text classification. In this method, the few-shot model utilised previously is adapted across language-related tasks by acquiring, transferring knowledge and interpreting the semantics of the labels in the source task and matches between these labels and the input text of the target task.

To this end, we have utilised ArSarcasm-v2 (Abu Farha et al. 2021), a benchmark Arabic dataset which encompasses a couple of classification tasks including: sentiment analysis and sarcasm detection. Despite being independent tasks, they both include similar components of language comprehension. Two experiments were conducted; the first one involved training the model on the sentiment task utilising SAtour manual and automatic sentiment labels to predict sarcasm labels from the ArSarcasm-v2 dataset. In the other experiment, we switched the tasks to predict sentiment labels from sarcastic text.

In our experiments, we utilised Google Colab as the computing environment, with Python version 3.10 and PyTorch version 0.18.1 + cu121. To ensure consistent and reproducible results, a random seed of 42 was set for all experimental runs. For the training of traditional Baseline models and Deep Learning Baseline models, we employed 10 epochs, while Zero/Few-shot Models were trained over 1000 episodes. No early stopping criteria were applied in any of the models. These standardised settings were maintained to guarantee a fair and rigorous comparison across the different models evaluated in this study.

6 | Discussion

We conducted all experiments in a three-way classification approach with all three classes. Table 5 lists the evaluation results of the baseline models on the test set. For evaluation, we reported Accuracy and F1-score, where accuracy is calculated as:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{6}$$

and $F1$-score as follows:

$$F1 = (TP) / \left(TP + \frac{1}{2}(FP + FN)\right) \tag{7}$$

From the previous experimental results, we can observe that MARBERT achieved the best performance on the test set of SAtour when manual annotations were considered. However, zero-shot annotations performed the best in predicting sarcasm classes in a few-shot cross-domain approach.

Moreover, the variety of dialects occurring in the dataset of Modern Standard Arabic (MSA) and Dialectal Arabic (DA) like the gulf dialect creates a challenge for the utilised classifiers. However, MARBERT was pre-trained on 128 GB of Arabic tweets comprising both MSA and diverse DA; this explains the superior performance of this model over the other utilised transformer-based models such as AraBERT.

It is worth mentioning that the few-shot model utilised in (Basabain et al. 2023) was adapted in this study to benchmark our collected data by applying the Zero/few-shot approach; however, we have changed the AraBERT text encoder with (sn-xlm-roberta-base-snli-mnli-anli-xnli). In a 10-way classification approach, their results show that AraBERT can produce better representations than the model in our experiment, with an accuracy of up to 0.833 for manual annotations.

TABLE 5 | Baseline models results on SAtour in a manual and automatic annotation of a three-way classification approach.

	Manual annotation		Zero-shot annotation with (*arabic_xlm_xnli*)	
Baseline model	**Acc**	***F*1-score**	**Acc**	***F*1-score**
TF-IDF + SVM	0.620	0.621	0.610	0.593
AraBERT	0.822	0.820	0.703	0.695
MARBERT	**0.867**	**0.856**	0.722	0.701
ArSarcasm-v2 (sarcasm) ➔ SAtour (sentiment)	0.436	0.421	0.612	0.596
SAtour (sentiment) ➔ ArSarcasm-v2 (sarcasm)	0.854	0.842	**0.841**	**0.845**
Features extracted utilising (*sn-xlm-roberta-base-snli-mnli-anli-xnli*) (with sentence transformers) + few-shot classification utilising Prototypical networks	0.623	0.596	0.524	0.516

Note: Bold values indicate the highest values.

TABLE 6 | Visualisation of AraBERT and MARBERT attention maps on SAtour with manual annotations. Higher attention weights are presented with darker highlights.

Model	Predicted label	Word importance
AraBERT	Negative	[CLS] ميسي كلب فلوس لدر ##جه يعلن عن اشياء تافهه مثل موسم جده ومو ##سم الرياض [SEP]
	Neutral	[CLS] تابع ##وني موسم جده [SEP]
	Positive	[CLS] بهر ##ا العالم وصر ##نا محط الانظار ابداع يوليف ##ارد رياض سيتي يوم الثن ##كر [SEP]
MARBERT	Negative	[CLS] ميسي كلب فلوس لدرجه يعلن عن اشياء تافهه مثل موسم جده وموسم الرياض [SEP]
	Neutral	[CLS] تابعوني موسم جده [SEP]
	Positive	[CLS] به ##ردا العالم وصردا محط الانظار ابداع يولي ##فارد رياض سيتي يوم التنك ##ر [SEP]

We can also explain the difference between the learning process of these two BERT-based models based on the attention weight of words in the input with respect to the [CLS] token. Table 6 demonstrates how each model differently tokenises the input and attends relevant words to the predicted class.

SAtour dataset was preliminarily cleaned as discussed in Section 3. However, the sarcastic tweets problem still exists, which may lead to a vague prediction and create a challenge for sentiment classifiers. Thus, future work should reconsider these tweets and represent them based on the implicit sarcasm beneath them. Overall, manual annotations helped for better sentiment classifications. Thus, further work is needed to enhance the ability of automatically labelling Arabic short-text using a zero-shot method.

In general, the performance scores in this study can be attributed to several key factors. MARBERT's pre-training on a large corpus of both Modern Standard Arabic and various dialects, such as Gulf dialect, enhanced its capacity to recognise minor differences in feeling expressions by enabling it to efficiently understand linguistic details and contextual meanings. The combination of manual and automatic annotations provided a diverse and rich training dataset, which improved the model's generalisation capabilities. Manual annotations offered high-quality, human-verified data, while the zero-shot labelling approach enabled the model to leverage pre-existing knowledge from different domains and apply it effectively to new, unseen data. Additionally, MARBERT's attention mechanism helped the model focus on relevant words and contextual cues, leading to more accurate predictions. The model's architecture, with multiple layers and attention heads, further enabled the capture of intricate patterns and relationships within the text, particularly suited to the rich morphology and syntax of the Arabic language. These factors—extensive pre-training, diverse annotation datasets, effective attention mechanisms and a sophisticated model architecture—collectively contributed to the high performance scores observed in our sentiment classification tasks.

7 | Conclusion

Twitter is an influential information sharing tool and a rich source of opinion texts on a variety of topics including business, economics, politics, society, and travel. In this study, we presented SAtour, a dataset for Arabic short text in the domain of tourism. The corpus was annotated for sentiment classification as either positive, negative, or neutral in both manual and automatic approaches. For the automatic labelling approach, we developed a zero-shot model to assign sentiment labels for the dataset unlabeled tweets. We then utilised both manual and automatic annotations in a zero-shot cross-domain classification approach to predict sentiment labels from a model trained on

sarcasm classes and vice versa. The zero-shot approach was employed not to enhance the results of other methods but to directly compare its performance with manual annotation. This methodology allowed us to evaluate the effectiveness of zero-shot learning as a standalone technique for annotating unlabeled data. Our findings demonstrated that zero-shot learning is a viable and promising alternative, achieving competitive performance levels. This highlights the potential of zero-shot learning for text annotation tasks, particularly in situations where manual labelling is not feasible.

Moreover, in this study, we developed a robust approach for sentiment analysis in the tourism domain, utilising zero-shot labelling and transformer-based models. A comparative study was carried out utilising different baseline models trained on both generated labels. Experimental results show that MARBERT exhibited the best performance in terms of the prediction accuracy rate, that is, it scored up to 87% in a three-way classification approach when training the models using manual annotations. However, zero-shot automatic labels scored 84% in models trained in a cross-domain approach to predict sarcasm classes from sentiment labels. Applying this approach to other domains requires several adjustments. Establishing domain-specific datasets is necessary to capture the unique characteristics and contextual complexities. Additionally, pre-training or fine-tuning the transformer-based models on these domain-specific corpora will enhance their ability to interpret and classify sentiments accurately. For instance, extending this method to healthcare, finance or customer service would require collecting relevant text data from these fields and incorporating additional domain-specific labels. By making these modifications, our solution can be effectively adapted to various domains, ensuring broad applicability and robustness.

Furthermore, the methodologies and techniques employed in this study, including the zero-shot labelling approach and the use of transformer-based models, have the potential to be adapted to other languages. The flexibility of transformer-based models allows for fine-tuning and pre-training on diverse linguistic datasets. By leveraging pre-trained multilingual models such as mBERT or XLM-R and adapting them with specific language corpora, our approach can be extended to handle sentiment analysis tasks in various languages. This adaptability underscores the potential of our approach to contribute significantly to sentiment analysis across different linguistic contexts beyond Arabic, accounting for the unique linguistic and contextual nuances of each language.

For future work, this corpus can be further utilised to overcome the challenges in Arabic text classification by considering the sarcasm beneath these tweets, topics, and intents. Moreover, different techniques can be applied to enhance features extracted from the text. For example, understanding the semantics of words in a tweet can affect the classification process.

Acknowledgements

Amir Hussain would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC)—Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1.

Data Availability Statement

The data that support the findings of this study are openly available in github at https://github.com/sbasabain/SAtour-A-dataset-with-Zero-shot-automatic-labelling-for-Arabic-short-text-Sentiment-Analysis.git.

Endnotes

[1] https://github.com/sbasabain/SAtour-A-dataset-with-Zero-shot-automatic-labelling-for-Arabic-short-text-Sentiment-Analysis.git.

[2] https://developer.twitter.com/en/docs/twitter-api.

[3] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa-did-madar-twitter5.

[4] https://huggingface.co/models.

[5] https://huggingface.co/morit/arabic_xlm_xnli.

[6] https://huggingface.co/symanto/sn-xlm-roberta-base-snli-mnli-anli-xnli.

References

Abdul-Mageed, M., A. Elmadany, and E. M. B. Nagoudi. 2021. "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1: Long Papers, 7088–7105. Association for Computational Linguistics.

Abu Farha, I., W. Zaghouani, and W. Magdy. 2021. "Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic." In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 296–305. Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.36.

Ainin, S., A. Feizollah, N. B. Anuar, and N. A. Abdullah. 2020. "Sentiment Analyses of Multilingual Tweets on Halal Tourism." *Tourism Management Perspectives* 34: 100658.

Al-Twairesh, N., H. Al-Khalifa, and A. Al-Salman. 2016. "AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 697–705. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1066.

Aldakhil, F. M. 2020. "Tourist Responses to Tourism Experiences in Saudi Arabia."

Alharbi, B. A., M. A. Mezher, and A. M. Barakeh. 2022. "Tourist Reviews Sentiment Classification Using Deep Learning Techniques: A Case Study in Saudi Arabia." *International Journal of Advanced Computer Science and Applications* 13, no. 6: 717–726.

Alkhaldi, R., D. Alsaffar, T. Alkhaldi, et al. 2022. "Sentiment Analysis for Cruises in Saudi Arabia on Social Media Platforms Using Machine Learning Algorithms." *Journal of Big Data* 9, no. 1: 1–28.

Almuqren, L., and A. Cristea. 2021. "AraCust: A Saudi Telecom Tweets Corpus for Sentiment Analysis." *PeerJ Computer Science* 7: e510.

Alomari, K. M., M. ElSherif, and K. Shaalan. 2017. *Arabic Tweets Sentimental Analysis Using Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-60042-0_66.

Alqarafi, A., A. Adeel, A. Hawalah, K. Swingler, and A. Hussain. 2018. *A Semi-Supervised Corpus Annotation for Saudi Sentiment Analysis Using Twitter. Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018*, 589–596. Springer.

Alsanad, A. 2018. "Arabic Topic Detection Using Discriminative Multi Nominal Naïve Bayes and Frequency Transforms." In *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 17–21. IEEE.

Annadani, Y., and S. Biswas. 2018. "Preserving Semantic Relations for Zero-Shot Learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7603–7612. IEEE.

Antoun, W., F. Baly, and H. Hajj. 2020. "AraBERT: Transformer-Based Model for Arabic Language Understanding." In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 9–15. European Language Resource Association.

Åslund, J. 2021. "Zero/Few-Shot Text Classification: A Study of Practical Aspects and Applications."

Basabain, S., E. Cambria, K. Alomar, and A. Hussain. 2023. "Enhancing Arabic-Text Feature Extraction Utilizing Label-Semantic Augmentation in Few/Zero-Shot Learning." *Expert Systems* 40, no. 8: e13329. https://doi.org/10.1111/exsy.13329.

Bojanowski, P. 2017. "Grave e Joulin a Mikolov t. Enriching Word Vectors With Subword Information." *ACL Anthology* 5: 135–146.

Brown, T., B. Mann, N. Ryder, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Chakravarthy, S., and J. Arutla. 2022. "Prognostication of Unseen Objects Using Zero-Shot Learning With a Complete Case Analysis." *Interdisciplinary Description of Complex Systems* 20: 454–468. https://doi.org/10.7906/indecs.20.4.10.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) Language Detection*, 4171–4186. Association for Computational Linguistics.

Dube, P., B. Bhattacharjee, S. Huo, P. Watson, B. Belgodere, and J. R. Kender. 2019. "Automatic Labeling of Data for Transfer Learning." *Nature* 192255: 122–129.

Elmadany, A., H. Mubarak, and W. Magdy. 2018. "Arsas: An Arabic Speech-Act and Sentiment Corpus of Tweets."

Fernández-Isabel, A., J. Cabezas, D. Moctezuma, and I. M. de Diego. 2023. "Improving Sentiment Classification Performance Through Coaching Architectures." *Cognitive Computation* 15, no. 3: 1065–1081.

Fleiss, J. L. 1971. "Measuring Nominal Scale Agreement Among Many Raters." *Psychological Bulletin* 76, no. 5: 378–382.

Flores-Ruiz, D., A. Elizondo-Salto, and M. Barroso-González. 2021. "Using Social Media in Tourist Sentiment Analysis: A Case Study of Andalusia During the COVID-19 Pandemic." *Sustainability* 13, no. 7: 3836.

Guellil, I., A. Adeel, F. Azouaou, and A. Hussain. 2018. "Sentialg: Automated Corpus Annotation for Algerian Sentiment Analysis." In *Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018*, 557–567. Springer.

Guellil, I., F. Azouaou, and M. Mendoza. 2019. "Arabic Sentiment Analysis: Studies, Resources, and Tools." *Social Network Analysis and Mining* 9, no. 1: 1–17.

Han, Z., Z. Fu, S. Chen, and J. Yang. 2022. "Semantic Contrastive Embedding for Generalized Zero-Shot Learning." *International Journal of Computer Vision* 130, no. 11: 2606–2622.

Kapadia, S. 2019. *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. Towards Data Science.

Kastrati, Z., A. S. Imran, and A. Kurti. 2020. "Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs." *IEEE Access* 8: 106799–106810.

Khalel, A., and M. El-Saban. 2018. "Automatic Pixelwise Object Labeling for Aerial Imagery Using Stacked u-Nets." *ArXiv*: 1803.04953.

Khan, Q., and H. N. Chua. 2021. "An Automated Topics Labelling Framework Using Zero-Shot Text Classification." *Journal of Engineering Science and Technology, Special Issue 6/2021 on ACSAT* 16: 46–59.

Kirilenko, A. P., S. O. Stepchenkova, H. Kim, and X. Li. 2018. "Automated Sentiment Analysis in Tourism: Comparison of Approaches." *Journal of Travel Research* 57, no. 8: 1012–1025.

Liu, H., D. Zhang, B. Yin, and X. Zhu. 2021. "Improving Pretrained Models for Zero-Shot Multi-Label Text Classification Through Reinforced Label Hierarchy Reasoning." *ArXiv*: 2104.01666, 2021.

Liu, X. 2020. "Analyzing the Impact of User-Generated Content on B2B Firms' Stock Performance: Big Data Analysis With Machine Learning Methods." *Industrial Marketing Management* 86: 30–39.

McCann, B., N. S. Keskar, C. Xiong, and R. Socher. 2018. "The Natural Language Decathlon: Multitask Learning as Question Answering." *ArXiv*: 1806.08730.

Mehrotra, R., S. Sanner, W. Buntine, and L. Xie. 2013. "Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling." In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 889–892. ACM.

Misargopoulos, A., F. Nikolopoulos-Gkamatsis, K. Nestorakis, et al. 2022. "Building a Knowledge-Intensive, Intent-Lean, Question Answering Chatbot in the Telecom Industry-Challenges and Solutions." In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 87–97. Springer International Publishing.

Mohammad, A.-S., M. M. Hammad, A. Sa'ad, A.-T. Saja, and E. Cambria. 2021. "Gated Recurrent Unit With Multilingual Universal Sentence Encoder for Arabic Aspect-Based Sentiment Analysis." *Knowledge-Based Systems* 227: 107540.

Mohammad, S., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. "SemEval-2018 Task 1: Affect in Tweets." In *Proceedings of the 12th International Workshop on Semantic Evaluation*, 1–17. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1001.

Nabil, M., M. Aly, and A. Atiya. 2015. "ASTD: Arabic Sentiment Tweets Dataset." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2515–2519. Association for Computational Linguistics.

Oueslati, O., E. Cambria, M. B. HajHmida, and H. Ounelli. 2020. "A Review of Sentiment Analysis Research in Arabic Language." *Future Generation Computer Systems* 112: 408–430. https://doi.org/10.1016/j.future.2020.05.034.

Pourpanah, F., M. Abdar, Y. Luo, et al. 2022. "A Review of Generalized Zero-Shot Learning Methods." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 4: 4051–4070.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1, no. 8: 9.

Ramanathan, V., and T. Meyyappan. 2019. "Twitter Text Mining for Sentiment Analysis on People's Feedback About Oman Tourism." In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, 1–5. IEEE.

Rezaei, M., and M. Shahidi. 2020. "Zero-Shot Learning and Its Applications From Autonomous Vehicles to COVID-19 Diagnosis: A Review." *Intelligence-Based Medicine* 3: 100005. https://doi.org/10.1016/j.ibmed.2020.100005.

Rosenthal, S., N. Farra, and P. Nakov. 2017. "SemEval-2017 Task 4: Sentiment Analysis in Twitter." In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518. Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2088.

Wang, H., X. Tang, Y.-H. Kuo, D. Kifer, and Z. Li. 2019. "A simple baseline for travel time estimation using large-scale trip data." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, no. 2: 1–22.

Wang, W., V. W. Zheng, H. Yu, and C. Miao. 2019. "A Survey of Zero-Shot Learning: Settings, Methods, and Applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, no. 2: 1–37.

Xie, G.-S., L. Liu, X. Jin, et al. 2019. "Attentive Region Embedding Network for Zero-Shot Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9384–9393. IEEE.

Xu, W., Y. Xian, J. Wang, B. Schiele, and Z. Akata. 2020. "Attribute Prototype Network for Zero-Shot Learning." *Advances in Neural Information Processing Systems* 33: 21969–21980.

Zhao, X., and J. Jiang. 2011. "An Empirical Comparison of Topics in Twitter and Traditional Media."

Zhao, X., S. Ouyang, Z. Yu, M. Wu, and L. Li. 2023. "Pre-Trained Languagemodels can be Fully Zero-Shot Learners." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, vol. 1, 15590–15606. Association for Computational Linguistics.

Zhong, R., K. Lee, Z. Zhang, and D. Klein. 2021. "Adapting Language Models for Zero-Shot Learning by Meta-Tuning on Dataset and Prompt Collections." *ArXiv*: 2104.04670.