IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 18 December 2023, accepted 22 January 2024, date of publication 8 February 2024, date of current version 26 February 2024.
Digital Object Identifier 10.1109/ACCESS.2024.3364367

RESEARCH ARTICLE

Bridging the Kuwaiti Dialect Gap in Natural Language Processing

FATEMAH HUSAIN [1], **HANA ALOSTAD** [2], **AND HALIMA OMAR** [3]
[1]Information Science Department, College of Life Sciences, Sabah AlSalem University City (Alshadadiya), Kuwait University, Safat 13060, Kuwait
[2]Computer Science Department, College of Arts and Sciences, Gulf University for Science and Technology, Hawally 32093, Kuwait
[3]Communication Disorders Science Department, College of Life Sciences, Sabah AlSalem University City (Alshadadiya), Kuwait University, Safat 13060, Kuwait
Corresponding author: Fatemah Husain (f.husain@ku.edu.kw)

ABSTRACT The available dialectal Arabic linguistic resources are very limited in their coverage of Arabic dialects, particularly the Kuwaiti dialect. This shortage of linguistic resources creates struggles for researchers in the Natural Language Processing (NLP) field and limits the development of advanced linguistic analytical and processing tools for the Kuwaiti dialect. Many other low-resource Arabic dialects are still not explored in research due to the challenges faced during the annotators' recruitment process for dataset labeling. This paper proposes a weak supervised classification system to solve the problem of recruiting human annotators called "q8SentiLabeler". In addition, we developed a large dataset consisting of over 16.6k posts serving sentiment analysis in the Kuwaiti dialect. This dataset covers several themes and timeframes to remove any bias that might affect its content. Furthermore, we evaluated our dataset using multiple traditional machine-learning classifiers and advanced deep-learning language models to test its performance. Results demonstrate the positive potential of "q8SentiLabeler" to replace human annotators with a 93% for pairwise percent agreement and 0.87 for Cohen's Kappa coefficient. Using the ARBERT model on our dataset, we achieved 89% accuracy in the system's performance.

INDEX TERMS Natural language processing, weak supervision, zero-shot language model, sentiment analysis, Arabic language, machine learning, Kuwaiti dialect.

I. INTRODUCTION

Arabic is widely spoken worldwide. The number of Arabic speakers exceeds 353.6 million.[1] Several countries and regions across Africa and Asia communicate in Arabic. Moreover, Arabic is the official language of the Islamic religion. Most of the Islamic literature and scripts are written in Arabic, making it a special language for Muslims worldwide. Despite these facts, the current state of research in Arabic Natural Language Processing (NLP) does not reflect the importance of the Arabic language. Multiple Arabic dialects, particularly the Kuwaiti dialect, are still not appropriately explored in NLP research. The preservation of Kuwaiti dialects and the enhancement of Kuwaiti culture will be promoted by addressing this gap in NLP research.

The Arabic language has multiple forms. The Classical Arabic Language (CAL) is the oldest form of Arabic (e.g., the holly Quran) [1], [2]. Modern Standard Arabic (MSA) is the official language for Arabic countries (e.g., media resources) [1], [3]. The last and most dominant form of Arabic is the Arabic dialects, which are the native language form of daily communication. The Arabic dialects have several variations and are not standardized using grammar rules [1]. In addition, Arabic users use Arabic dialects when writing text in online conversations, which makes the development of textual tools and analysis systems very challenging, hindering the accurate automatic process of Arabic social media content.

The Kuwaiti dialect is a unique Arabic dialect that has several distinguishing features. While researchers classify Arabic dialects into seven main categories: Egyptian, Levantine, Maltese, North African, Iraqi, Yemeni, and Gulf [1], each country in the Arabian Gulf has its specific dialect and in some cases sub-dialects. The scope of this study covers the Kuwaiti dialect.

This study aims to create a large Kuwaiti dataset for sentiment analysis text classification tasks while reducing

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

[1]https://www.worlddata.info/languages/arabic.php

the efforts and resources required in human annotation. The main contributions of this study can be summarized in the following points:

1) Creating a gold-labeled dataset using human annotators for sentiment analysis including 1,534 posts.
2) Developing the "Q8SentiLabeler", a weakly supervised learning system to label large datasets in Kuwaiti dialect for sentiment analysis.
3) Automatically constructing a comprehensive sentiment analysis Kuwaiti X (previously known as Twitter) dataset consisting of 16,667 posts (previously known as Tweets).
4) Creating lexicon lists to analyze the variations of linguistic features among different classes.
5) Evaluating the proposed "Q8SentiLabeler" system using the gold-labeled dataset.
6) Evaluating the dataset and the overall system performance using simple traditional machine learning classifiers and advanced deep learning language models.

The paper starts with general background information about the main topic, covering an introduction to the Kuwaiti dialect and the previous NLP studies on the Kuwaiti dialect, in addition to a revision on the available methods for data annotation and sentiment analysis. The Related Work section is followed by the Methodology section, which discusses in detail the procedures conducted to arrive at our proposed system, consisting of corpus development, exploratory data analysis, classification system construction, and performance evaluation. Then, the results and error analysis based on the results are discussed in detail in the Result section. The paper ends with a conclusion.

II. RELATED WORK

A. KUWAITI DIALECT

The state of Kuwait is located in the Arabian Peninsula. Historically, Kuwait has four sectors; Sharq (East), Qibla (West), Hay al−Wasaṭ (Middle Neighbourhood), and al−Mirqab (South) [4]. This geographic variation, in addition to some other demographic and cultural factors (e.g., profession, economic stability), leads to differences within the Kuwaiti dialects, creating multiple Kuwaiti sub-dialects. In some cases, distinguishing one Kuwaiti sub-dialect from others can be easily captured through minor variations in its phonetics sounds, such as the word sugar in MSA "سُكر / Sukar", which is pronounced as "شِكر / Shikar" by Kuwaitis from Sharq and "شُكر / Shukar" by Kuwaitis from Qibla.[2] In more extreme situations, different words are used to refer to the same meaning based on the geographic origin, for example, the word "سرير / sareer" is used by Kuwaitis from Qibla while the word "كِرفاية / kirfaia" is used by Kuwaitis from Sharq to refer to a bed in MSA "سرِير /sareer".

Kuwaitis have been exposed to continuous contact with several cultures, Arabic dialects, and languages; such as Cairene Arabic (i.e. Egyptian), dialects of Saudi Arabia, Turkish, Hindi, and Persian [4]. For example, the bicycle in Kuwait is called "قاري / Gari" borrowed from Hindi origin in which gari is used to describe a vehicle.[3]

Furthermore, Kuwait was a protectorate of the British Empire for 62 years, which also created an effect on the Kuwaiti dialect [5]. Multiple English words integrated into the Kuwaiti dialect with its identical English meaning such as "واير / Wire", and "هرن / Horn". "بنسل / Pencil".

In addition, there are some new phenomena recently experienced when communicating with the new generations in Kuwaiti societies. According to [5], young Kuwaitis are integrating the English language significantly into their dialect when communicating, which creates two phenomena: the "McChickens" and "Chicken Nuggets". The "McChickens" is a term used to describe bilingual young Kuwaitis who switch to English and use some English words while speaking in Kuwaiti dialect, while the "Chicken Nuggets" describes young Kuwaitis who are communicating in English because their Kuwaiti dialect is significantly weak [5].

This complex structure of the Kuwaiti dialect makes it very difficult to create a linguistic system that can automatically process Kuwaiti text accurately.

B. NATURAL LANGUAGE PROCESSING AND THE KUWAITI DIALECT

The NLP field is in continuous progress, with linguistic tools and applications enabling the creation of advanced solutions to the most challenging and complex linguistic problems. The ChatGPT[4] is a great example of how NLP can solve problems in a way that might exceed human abilities.

From a Kuwaiti dialect perspective, the available linguistic computational tools and resources are very limited and do not support the current advanced state of research in the NLP field. We use multiple search engines (e.g., Google, Yahoo, Bing) and research databases and libraries online (e.g., the Association for Computing Machinery Digital Library (ACM DL), the Institute of Electrical and Electronics Engineers (IEEE) Xplore, the Association for Computational Linguistics (ACL)) to search for publications in computational linguistic or NLP for the Kuwaiti dialect. After extensive searching, we were able to find only three publications dedicated to the Kuwaiti dialect.

The first study was published in 2014 in which the authors performed a linguistic analysis to develop a Kuwaiti sentiment classification system [6]. Reference [6] create multiple keyword lists of different sentiments and emotions based on a large dataset collected from X. However, their dataset is not publicly available and the keywords are very general Arabic terms, which are used in other Arabic dialects as well. Moreover, in the study conducted a long time ago,

[2] https://www.alanba.com.kw/ar/kuwait-news/620552/23-01-2016

[3] https://www.kuwaittimes.com/origins-kuwaiti-dialect/

[4] https://openai.com/blog/chatgpt/

the classification systems applied in their experiments are basic machine learning classifiers that do not use advanced language models commonly used in the current NLP field.

The second study was published in the proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (WANLP) in 2022. The authors propose a weak supervised system to support automatic data labeling for a large Kuwaiti sentiment analysis corpus from X [7]. They cover several social events that influence Kuwaiti society beyond the X sphere and that span one year of time frame. They applied traditional machine learning classifiers as well as advanced language models [7]. Despite that, their dataset is not publicly available.

The third study was also published in the proceedings of the 7th WANLP in 2022. This paper consists of a pilot study for a gender classification dataset for the Kuwaiti dialect from a chatting WhatsApp for a reading group [8]. This study provides analytical linguistic variations between men and women, which include text and pictograms (e.g., emoji) [8]. However, this dataset is dominated by female participants (28 women), and male participants are a minority (14 men), which might not accurately serve the study's main goal. In addition, it focuses on only one chat group with a particular theme. The dataset is not evaluated using a classification model to assess its performance for the gender classification task.

Based on our discussion in this section, the research gap in the Kuwaiti dialect is very significant. Multiple NLP tasks are not explored such as question answering, speech recognition, text generation, and text summarization.

C. DATA ANNOTATION METHODS

Data annotation consists of providing labels/classes to data based on the study's goal. The annotation could be on the word level or the sentence level. This process is very helpful as most online text is available in an unstructured format; thus, providing labels can help machines understand the data and identify important elements. Moreover, data annotation is one of the most challenging phases in the text classification system pipeline. The following are some of the challenges accompanied by recruiting human annotators in the labeling process:

- Advanced deep learning and transfer learning algorithms require very large size labeled datasets, which might be time-consuming and not practical to create.
- Subject Matter Experts (SMEs) have limited time; thus, it is difficult for them to provide labels for a large dataset.
- In the case of labeling through crowd-sourcing, the labeling task will be very costly and raise some quality issues (e.g., proficiency in the subject, personal and racial bias, background knowledge effects, and agreement among annotators).
- Data privacy concerns impact the annotation process and the recruitment of annotators.
- Difficulties to Arabic annotators and in Arabic data as it has multiple forms and dialects, which creates inconsistency in meaning for some phrases and lexicons [9].

Alternative approaches are currently available to label/annotate data, either with or without help from SMEs. The following is a description of each approach:

1) **Active learning**: is a paradigm in machine learning that aims to select the most informative data points to label from a data stream. The hypothesis in active learning is that if the learning algorithm can choose the data it learns from, it will perform better with less training. The active learner aims to achieve high accuracy using as few labeled instances as possible, thereby, minimizing the cost of obtaining labeled data [10].
2) **Semi-supervised learning**: is a machine-learning approach that combines both labeled and unlabeled data for training models. It aims to leverage the information present in the unlabeled data to improve the performance of the model [11].
   Weak-supervised learning is a type of semi-supervised learning, [12] defines it as a collection of techniques in machine learning in which models are trained using labeling functions as sources of incomplete, inexact information that are easier to provide than hand-labeled data. The noisy, weak labels are combined using a generative model trained based on the accuracy of labeling functions; the accuracy is derived from agreement and disagreement of the labeling functions and used to form the training data. Snorkel [13] is an example of a weak supervised learning framework proposed by researchers at Stanford AI Lab. Snorkle was able to build models 2.8x faster than human-labelers with 45.5% better predictive performance on average [13].
3) **Transfer Learning**: is a machine learning technique that has gained significant attention due to its ability to leverage knowledge from a source domain to improve learning in a target domain with limited labeled data [14].
   Zero-Shot (ZS) learning is related to transfer learning, based on [15], the ZS model can predict the class of the unlabeled sample, even if the model is not trained on those classes. It relies on the model's ability to transfer knowledge. Using Natural Language Inference (NLI), ZS learning is suitable in a setting where no labeled data is provided. The ZS models leverage the semantic similarity between labels and the text context [16]. In this type of experiment setup, the text to be labeled is treated as the premise, and the prompt hypothesis template is formed as "this example is about {label}". In addition, a set of expected labels is included in the training configuration. Finally, the entailment score tells us whether the promise is about that topic/label or not.

A good candidate to perform ZS learning in languages other than English is the XLM-RoBERTA (XLM-R) model. This model is trained on one hundred languages, including Arabic and many other low-resource languages [17]. Another good candidate model, is the Multilingual mDeBERTa, which is based on [18]. It is the best-performing multilingual base-sized State-Of-The-Art (SOTA) model in the cross-lingual Natural Language Inference (XNLI).

D. SENTIMENT ANALYSIS

The process of understanding users' attitudes and emotions is called sentiment analysis [19]. It has two main measurements. The first measurement is polarity, which gives a measurement of whether the statement is positive, negative, or neutral. Polarity is based on a floating number score within the range of -1.0 and 1.0. Accordingly, positive sentiment has a value close to 1, while negative sentiment has a value close to -1. The second measurement is subjectivity, which gives a measurement of personal opinion content versus factual information, and it has a floating value within the range of 0.0 and 1.0. Statements closer to 0.0 are more objective than statements with a value closer to 1.0.

In addition, sentiment analysis can have fine-grained classification, as multiple classes, rather than just two or three, can be used for the sentiment classification [20]. It can also apply simple methods like lexicon-based methods and more advanced methods like deep learning language models for sentence prediction [21].

Sentiment analysis systems have several uses and applications. From business and commerce perspectives, sentiment analysis is used to understand customer behavior, satisfaction, and feedback [22], [23]. In journalism, identifying the reliability of news articles is supported by sentiment analysis methods [24]. In the stock market, sentiment analysis helps investors to make timely and informed trading decisions [25]. Moreover, in social media platforms, sentiment analysis is commonly used to support content moderation [26].

The available research in sentiment analysis for the Kuwaiti dialect is very limited. As we mentioned earlier, the first sentiment classification for the Kuwaiti dialect was conducted by [6] and the second by [7]. Similarly, recent studies covering Arabian Gulf dialects for sentiment analysis developed several datasets and classification systems. For the Bahraini dialect, a parallel balanced dataset of English, MSA, and Bahrani dialects consisting of 5,000 product reviews and a dataset of 500 movie comments in Bahraini dialect were created for a sentiment analysis system [27]. In [28], the authors introduced the first Emirati sentiment analysis dataset, which consists of 70,000 Instagram comments. The Saudi dialect is the most studied Gulf dialect, several sentiment analysis resources were created including but not limited to the followings: (1) [29] develop 2010 posts dataset for sentiment analysis; (2) [30] collect 32,063 Saudi posts; (3) [31] construct a dataset of 11,764 posts about Saudi universities; (4) in [32], the authors create a dataset of 22,433 reviews of tourist places.

From our research, we find that although sentiment analysis systems are in continuous research, not many studies have focused on applying semi-supervised methods to conduct it.

III. METHODOLOGY

A. CORPUS DEVELOPMENT

The dataset used in this study is collected over a time frame of one year. Having a large time frame ensures the diversity of data content in terms of themes and contributors, and removes any bias that might occur within the society and online users regarding a particular event. Accordingly, four social events covering four general themes; feminism, justice, humanity, and religion; were selected. Data consists of raw text of posts from X.

1) POSTS EXTRACTION

Public posts were collected from X developer API. We select all Arabic text posts and posts that contain equal to or more than three tokens/words. The chosen four controversial events all happened in Kuwait, impacting the Kuwaiti mainstream. Followings are a brief description of each event along with the hashtags used to extract posts:

- Feminism event. These posts were collected in April 2021. Farah Akbar is a Kuwaiti woman who was threatened and harassed, then killed wildly by a man after she rejected his marriage proposal. Two hashtags were selected to extract these posts; #عزاء_النساء and #جريمة_قتل_صباح_السالم.
- Justice event. These posts were collected in October 2021. Dalal Al-Abd Al-Jader is a Kuwaiti girl who was killed by her mother and kept for five years inside the apartment without being buried. The hashtag used to extract these posts is #العدالة_لدلال_العبدالجادر.
- Humanity event. Bideon, Bidoon, or Bedun refer to a stateless Arab minority in Kuwait. Bideon struggles in their life as the law prevents them from having basic life qualities, such as nationalities, civil identification cards, and marriage contracts, among multiple other official documents, which causes difficulties in finding jobs, receiving healthcare service, and education. Posts shared in February 2022 about Bidoon were selected because they coincided with the Moroccan child Rayan's incident, which received the attention of an overwhelming worldwide number of online users including Kuwaitis. This attitude by Kuwaitis toward Rayan's incident created an extreme wave of anger among the Bidoon community. As a result of that, they protest in Kuwaiti streets urging for their citizenship and their civil rights. Three hashtags were used to extract these posts including #البدون #البدون_اولويه #الطفل_البدون_عبدالعزيز.
- Religious event. Posts related to this event were extracted in April 2022. Sheikh Al-Hazeem is a Kuwaiti

FIGURE 1. **A sample of positive posts.**

Shia clergy who was attacked while in the mosque by government officials trying to take Zakat (i.e. donation) money collected from people. The hashtag used to collect these posts is #محشوم_الشيخ_مهدي_الهزيم.

2) DATA CLEANING AND FILTERING

During data cleaning, we first removed posts that had been reposted, hashtags "#", and user mentions "@", then all duplicated posts were removed. We do not perform in-depth text cleaning during the corpus development phase as we plan to adopt different levels of text preprocessing based on the classification models conducted in the proceeding phases.

3) POSTS ANNOTATION

The annotation process depends on the expressed feeling in the post's text. Thus, positive posts demonstrate feelings such as happiness, fun, and pride, while negative posts show feelings such as sadness and contempt. Furthermore, posts that do not contain clear emotional expression are categorized as neutral. Sample posts from the dataset for each label are presented in figures 1, 2, and 3. We adopt two approaches during the annotation step, the first approach is the traditional approach in which human annotators are recruited to manually label the data, and the second approach consists of our proposed weak supervised labeling system.

1) **Human Labeling** We recruited seven Kuwaiti annotators between the ages of 17 and 24 years who are proficient speakers and writers in the Kuwaiti dialect to manually annotate a set of 2,100 posts (300 posts per annotator). All annotators were given detailed annotation instructions that consisted of class definitions and samples from each class. In addition, annotators were asked to participate in a background survey to collect demographic information that might impact the annotation process, and a pilot study was performed to accurately guide the annotators during the annotation step. The pilot study includes 15 posts; 5 posts were provided with their labeled as samples from different labels and 10 posts without their labels were provided to test the annotators. Thus, annotators who correctly annotated the ten testing

FIGURE 2. **A sample of neutral posts.**

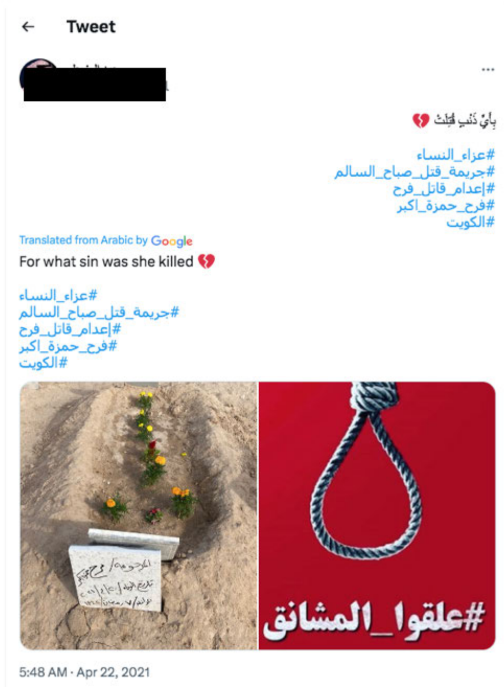

FIGURE 3. **A sample of negative posts.**

posts proceeded further in the annotation process, with 300 more posts to be labeled. An expert annotator also examined these labeled posts for their accuracy and correctness. As a result, this set of 2,100 was filtered to only include 1,534 posts, which we called the gold-labeled dataset. This set of posts was used to further examine our approach to data labeling using weak supervision techniques as we will illustrate in the next sections.

2) **Q8SentiLabeler System Architecture** Figure 4 illustrates the architecture for the Q8SentiLabeler system; the following is a description of each step we followed in our experiment setup:

a) For each experiment, we selected one of the following ten prompt phrases to use as the first input to the ZS model labeling function:
- The sentiment of this post is { }
- The tweet (post) sentiment is { }
- The sentiment in the post is { }
- The label of the tweet (post) sentiment is { }
- The sentiment label of this post is { }
- What is the sentiment label of this post is { }
- What is the sentiment of this post is { }
- What is the sentiment classification of this post { }
- Tell me what is the sentiment classification label of this post { }
- What is the sentiment classification label of this post { }

b) We fixed the labels to positive, negative, and neutral, and then, we used them as the second input to the ZH model labeling function, the same as in [7].

c) Using the Snorkel framework, we applied three ZS model labeling functions on the unlabeled posts dataset.

d) We trained Snorkel **LabelModel** based on the labels resulting from the labeling functions; as a result, each labeling function will get a weight based on its agreement and disagreement level with other labeling functions.

e) We used the trained model resulting from the previous step to predict the labels for the unlabeled posts dataset.

f) We used the gold-labeled dataset to validate the trained labeling model and evaluate the quality of labeling by applying the labeling functions on the human-annotated (gold-labeled) dataset and then using the trained model to predict the labels.

g) To evaluate the performance of Q8SentiLabeler, we calculated the accuracy and macro-F1 values.

h) To evaluate the quality of Q8SentiLabeler labels compared to human labelers, we calculated annotation agreement and the Cohen Kappa score.

Furthermore, we conducted experiments to test the effect of combining all or partial prompts as inputs to ZS labeling functions in one experiment.

TABLE 1. Dataset distribution based on sentiment classes and events.

	Events				Totals
	Feminism	Justice	Humanity	Religious	
Positive	534	902	1,089	5,377	**7,902**
Negative	832	1,966	1,465	3,642	**7,905**
Neutral	61	99	183	517	**860**
Totals	**1,427**	**2,967**	**2,737**	**9,536**	**16,667**

B. EXPLORATORY DATA ANALYSIS

To better understand the content of the dataset, we conducted some exploratory data analysis techniques to investigate the variations among posts from different classes. Table 1 reports the distribution of posts based on their classes per event. As can be noticed, the largest proportion of the data is related to the last event, a religious event, and the smallest one belongs to the first event, feminism. This might impact the overall theme of the sentiment toward political and religious issues.

We further investigate posts' content based on their classes by extracting collocations. Collocations refer to terms of a sequence of two or more words based on a common or traditional way of saying them [33]. In our experiments, we apply the same collocation extraction approach defined by [34] in which they adopt five methods without applying part-of-speech pattern filtering. The five methods are the frequency-based, the t-test, the chi-square test, the likelihood ratio, and the Pairwise Mutual Information (PMI). Each method applies some statistical computational measures that differentiate the outcome from them. We first classify the dataset based on their classes and then we extract collocations using the five methods. We consider collocations consisting of two words (bi-grams) and three words only (tri-grams). Tables 2 to 7 show the results from each method per class in detail. For the bi-grams lists, we selected the top 15 collocations. For the tri-grams lists, we selected the top 10 collocations, except for the negative posts for which we couldn't generate a large number of collocations, so we selected the top 5 collocations.

C. CLASSIFICATION SYSTEM CONSTRUCTION

We randomly partition the dataset into three subsets. Firstly, the train set with 60% of the total number of posts. Secondly, the validation set with 20%. The last set is the test set consisting of 20% of the total posts. Moreover, we use the "train_test_split" function from the "scikit-learn" Python package to ensure randomness and equal label distributions. The distribution of each set is as follows:
- **Train set**: consisting of 9,224 posts; positive: 4,404 posts, neural: 463 posts, and negative: 4,357 posts.
- **Validation set**: consisting of 3,233 posts; positive: 1,576 posts, neural: 190 posts, and negative: 1,567 posts.
- **Test set**: consisting of 3,334 posts; positive: 1,560 posts, neural: 174 posts, and negative: 1600 posts.

We start by training the classification models using the train set, then we initially evaluate the performance of the models using the validation set. After that, we merge

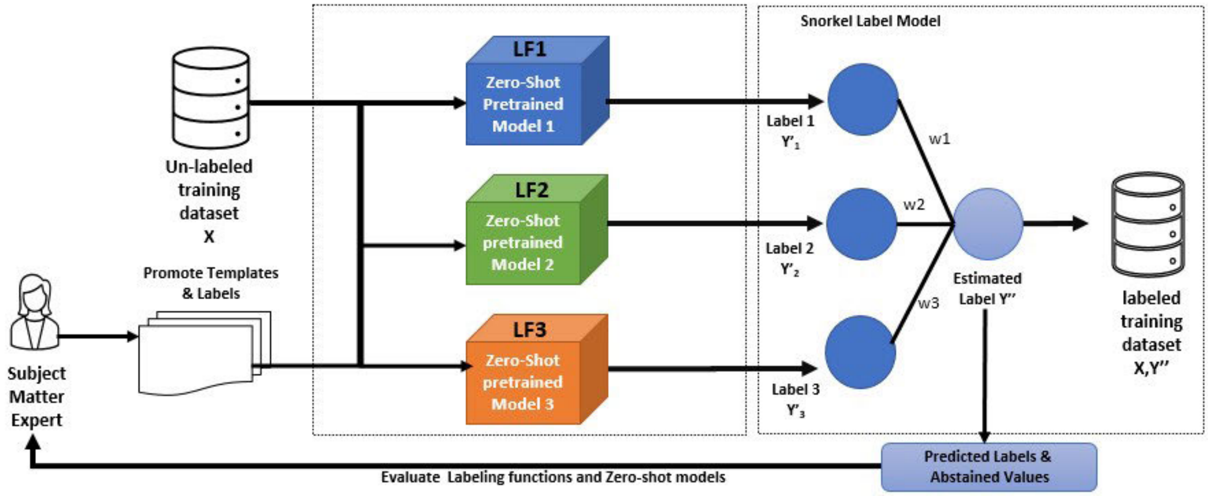

FIGURE 4. Q8SentiLabeler system architecture.

TABLE 2. Bi-grams for the positive posts.

#	Frequency	PMI	T-Test	Chi-Square	Likelihood
1	مرزوق الغانم Marzouq Al-Ghanim	مجلس الوزراء council of ministers	صلاة العيد Eid prayer	اتصالا هاتفيا Phone call	نواف الأحمد Nawaf Al-Ahmad
2	عيد الفطر Eid Al-Fitr	خفر السواحل coastguard	تقبل الله may Allah accept it	سلطنة عمان Sultanate of Oman	الملك سلمان King Salman
3	صلاة العيد Eid prayer	صندوق النقد Monetary Fund	عيد الفطر Eid Al-Fitr	خادم الحرمين custodian of the two holy mosques	الحشد الشعبي The Popular Mobilization Forces
4	عيدكم مبارك Blessed Eid	تواصل معنا connect with us	ولي العهد crown prince	خفر السواحل coastguard	صلاة العيد Eid prayer
5	مجلس الأمة National assembly	الإمام الحسين Al-Imam Al-Hussain	أمير البلاد prince of the country	ساحة الإرادة Al-Erada Square	تقبل الله may Allah accept it
6	أمير البلاد Prince of the country	خادم الحرمين Custodian of the two holy mosques	وزارة الداخلية Ministry of interior	ولي العهد Crown prince	الإمام الحسين Al-Imam Al-Hussain
7	مجلس الوزراء Council of ministers	الملك سلمان King Salman	البدون أولوية Albidoon are priority	جورج قرداحي George Kordahi	وزارة الداخلية Ministry of interior
8	وزارة الداخلية Ministry of inferior	جريمة قتل Murder	نواف الأحمد Nawaf Al-Ahmad	صندوق النقد Monetary fund	خفر السواحل Coastguard
9	قاتل فرح Farah's murderer	سلطنة عمان Sultanate of Oman	الحشد الشعبي The popular mobilization forces	تواصل معنا Connect with us	وزير الدفاع Minister of defense
10	خادم الحرمين Custodian of the two holy mosques	وزارة الإعلام Ministry of information	عيدكم مبارك Blessed Eid	مجلس الوزراء Council of ministers	جريمة قتل murder
11	أول حب First love	وزير الدفاع Minister of Defense	قاتل فرح Farah's murderer	جريمة قتل murder	أول حب first love
12	الحرمين الشريفين The two holy mosques	ولي العهد Crown prince	صلاة العيد Eid prayer	وزير الدفاع Minister of Defense	جريمة قتل National Assembly
13	الملك سلمان King Salman	أمير البلاد Prince of the country	شهر رمضان Month of Ramadan	نواف الأحمد Nawaf Al-Ahmad	عيد الفطر Eid Al-Fitr
14	الإمام الحسين Al-Imam Al-Hussain	ساحة الإرادة Al-Erada Square	مجلس الأمة National assembly	أمير البلاد prince of the country	جورج قرداحي George Kordahi
15	ولي العهد Crown prince	اتصالا هاتفيا Phone call	الملك سلمان King Salman	مرزوق الغانم Marzouq Al-Ghanim	مجلس الوزراء council of ministers

both, the validation set and the training set, to train the classification models and evaluate them using the test set.

Two main categories of classification models were applied; we used simple baseline models and deep learning language models as explained in the next section.

1) BASELINE MODELS

We applied four baseline classification models including; Logistic Regression (LR), Support Vector Machine (SVM), Multinomial Naive Bayes (M-NB), and Bagging.

For the baseline models, we performed some additional preprocessing techniques to the basic data filtering and cleaning techniques performed earlier. Some Arabic letters are written in more than one form based on the location of the letter within the word. Accordingly, we normalize some letters including Alif (أ ، آ ، إ to ا), Alif Maqsura (ي ، ئ to ى), and Ta Marbouta (ة to ه). Some words contain repetitions in letters more than two times, which were reduced to two times only. Hashtags were further preprocessed, after removing the "#" symbol, we replaced the "_" symbol with a space. In addition, text was cleaned to remove numbers, kashida (create justification by elongating characters), HTML tags, more than one consecutive space, diacritics, punctuation, and Arabic stop-words.[5]

All models were applied with a 2-5 characters-based TF-IDF vectorizer as previous studies demonstrate the importance of applying a character-based feature when using a user-generated content dataset [35], [36]. User-generated content contains misspelling errors and obfuscating terms, which made character-based techniques the best choice during feature extraction as it is a language-independent approach.

2) DEEP LEARNING LANGUAGE MODELS

In addition to the baseline models, multiple deep-learning language models based on the Bidirectional Encoder

[5]https://github.com/mohataher/arabic-stop-words

TABLE 3. Bi-grams for the neutral posts.

#	Frequency	PMI	T-Test	Chi-Square	Likelihood
1	الملك سلمان King Salman	أمير البلاد Prince of the country	شهر رمضان Ramadan month	نواف الأحمد Nawaf Al-Ahmad	عيد الفطر Eid Al-Fitr
2	أول حب First love	وزير الدفاع Minister of defense	قاتل فرح Farah's murderer	جريمة قتل Murder	أول حب First love
3	مجلس الوزراء Council of ministers	الملك سلمان King Salman	البدون أولوية Albidoon are priority	جورج قرداحي George Kordahi	وزارة الداخلية Ministry of interior
4	الحرمين الشريفين The two holy mosques	ولي العهد Crown prince	صلاة العيد Eid prayer	وزير الدفاع Minister of defense	جريمة قتل National assembly
5	مرزوق الغانم Marzouq Al-Ghanim	مجلس الوزراء council of ministers	صلاة العيد Eid prayer	اتصالا هاتفيا call	نواف الأحمد Nawaf Al-Ahmad
6	قاتل فرح Farah's murderer	سلطنة عمان Sultanate of Oman	الحشد الشعبي Popular Mobilization Forces	تواصل معنا Connect with us	وزير الدفاع Minister of defense
7	ولي العهد Crown prince	اتصالا هاتفيا Phone call	الملك سلمان king Salman	مرزوق الغانم Marzouq Al-Ghanim	مجلس الوزراء council of ministers
8	عيدكم مبارك Blessed Eid	تواصل معنا Connect with us	ولي العهد Crown prince	خفر السواحل Coastguard	صلاة العيد Eid prayer
9	مجلس الأمة National assembly	الإمام الحسين Al-Imam Al-Hussain	أمير البلاد Prince of the country	ساحة الإرادة Al-Erada Square	تقبل الله May Allah accept it
10	عيد الفطر Eid Al-Fitr	خفر السواحل Coastguard	تقبل الله May Allah accept it	سلطنة عمان Sultanate of Oman	الملك سلمان King Salman
11	وزارة الداخلية Ministry of inferior	جريمة قتل murder	نواف الأحمد Nawaf Al-Ahmad	صندوق النقد Monetary fund	خفر السواحل Coastguard
12	خادم الحرمين Custodian of the two holy mosques	وزارة الإعلام Ministry of information	عيدكم مبارك Blessed Eid	مجلس الوزراء Council of ministers	جريمة قتل Murder
13	الإمام الحسين Al-Imam Al-Hussain	ساحة الإرادة Al-Erada Square	مجلس الأمة the national assembly	أمير البلاد prince of the country	جورج قرداحي George Kordahi
14	صلاة العيد Eid prayer	صندوق النقد Monetary Fund	عيد الفطر Eid Al-Fitr	خادم الحرمين Custodian of the two holy mosques	الحشد الشعبي The Popular Mobilization Forces
15	أمير البلاد Prince of the country	خادم الحرمين Custodian of the two holy mosques	وزارة الداخلية Ministry of interior	ولي العهد Crown prince	الإمام الحسين Al-Imam Al-Hussain

TABLE 4. Bi-grams for the negative posts.

#	Frequency	PMI	T-Test	Chi-Square	Likelihood
1	وزارة الداخلية Ministry of Interior	وزير الداخلية Minister of Interior	الجهاز المركزي Central agency	الضحية القادمة The next victim	ساحة الإرادة Al-Erada Square
2	الإمام الحسين Al-Imam Al-Hussain	عيدكم مبارك Blessed Eid	قضية البدون Albidoon case	عيد الفطر Eid Al-Fitr	الحشد الشعبي Popular mobilization forces
3	الحشد الشعبي Popular mobilization forces	ساحة الإرادة Al-Erada Square	وزارة الداخلية Ministry of Interior	ساحة الإرادة Al-Erada Square	قاتل فرح Farah's murderer
4	مجلس الأمة National assembly	الضحية القادمة The next victim	وزارة الإعلام Ministry of Information	عبيد الوسمي Obaid Alwasmi	سعد العبدالله Saad Al-Abdullah
5	عيد الفطر Eid Al-Fitr	الله يرحمها May Allah have mercy on her	البدون أولوية Albidoon are priority	كفى استهتار Enough recklessness	الجهاز المركزي Central agency
6	ساحة الإرادة Al-Erada Square	الإمام الحسين Al-Imam Al-Hussain	عبيد الوسمي Obaid Alwasmi	مطلب شعبي Popular demand	كفى استهتار Enough recklessness
7	وزارة الإعلام Ministry of Information	مجلس الوزراء Council of ministers	قاتل فرح Farah's murderer	سعد العبدالله Saad Al-Abdullah	وزارة الداخلية Ministry of interior
8	عيدكم مبارك Blessed Eid	كفى استهتار Enough recklessness	الضحية القادمة The next victim	شارع الهرم Al-Haram street	وزير الداخلية Minister of Interior
9	البدون أولوية Albidoon are priority	مجلس الأمة National Assembly	مجلس الوزراء Council of ministers	جورج قرداحي George Kordahi	الضحية القادمة The next victim
10	الله يرحمها May Allah have mercy on her	زكاة الفطر Zakat Al-Fitr	ساحة الإرادة Al-Erada Square	صحة الطلاب Students' health	مجلس الأمة National Assembly
11	الضحية القادمة The next victim	الحشد الشعبي The Popular Mobilization Forces	مجلس الأمة National assembly	الطلاب برقبتكم Students are your responsibility	جريمة قتل Murder
12	وزير الداخلية Minister of Interior	عبيد الوسمي Obaid Alwasmi	فرح حمزة Farah Hamza	نشكو إليك We complain to you	عيد الفطر Eid Al-Fitr
13	مجلس الوزراء Council of ministers	عيد الفطر Eid Al-Fitr	كفى استهتار Enough recklessness	عبدالكريم الكندري Abdulkareem Al-Kandri	الإمام الحسين Al-Imam Al-Hussain
14	زكاة الفطر Zakat Al-Fitr	البدون أولوية Albidoon are priority	الحشد العبي Popular mobilization forces	حميد القلاف Humaid Al-Qallaf	مطلب شعبي Popular demand
15	فرح حمزة Farah Hamza	وزارة الإعلام Ministry of Information	عيد الفطر Eid Al-Fitr	علاء الحيدري Alaa Al-Haidari	البدون أولوية Albidoon are priority

Representations from Transformers (BERT) architecture were also adopted to construct our sentiment analysis system. The BERT-based language models used pre-trained language representations to downstream text classification tasks through a fine-tuning module. This fine-tuning module is also called transfer learning. During the fine-tuning process, pre-trained language representations are constructed using a neural network model for a known classification task. After that, fine-tuning is performed to use the same model for a new purpose-specific classification task, for our case it is used for sentiment analysis [37].

We adopted the following BERT models in our experiments:

1) The AraBERT Model [38]. It is a monolingual Arabic BERT model. It has various versions with variations in the model architecture and training corpus. In this study, ''bert-base-arabertv02-twitter'' is applied, which is trained by continuing the pre-training process using the masked language model pipeline with around 60 million Arabic posts. This version of AraBERT includes emoji in its vocabulary.[6]
2) The ARBERT Model [39]. It uses the same architecture of the BERT base model with a large MSA dataset that has been collected from 6 various sources.[7]
3) The MARBERT Model [39]. This model has been developed by the same authors as ARBERT, however, it was developed using a larger dialectal dataset than ARBERT with more tokens that are collected from randomly selected posts. It has the same architecture as ARBERT, but without the Next Sentence Prediction (NSP) objective as posts are concise and short.
4) The Microsoft Multilingual Model (MiniLM) [40]. It is a small and fast pre-trained model for language understanding and generation. It is distilled from the ''XLM-RoBERTa'' model, however, the transformer

[6]https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter
[7]https://github.com/UBC-NLP/marbert

TABLE 5. Tri-grams for the positive posts.

#	Frequency	PMI	T-Test	Chi-Square	Likelihood
1	عيد الفطر المبارك Blessed Eid Fitr	خادم الحرمين الشريفين Custodian of the tow holy mosques	تقبل الله طاعتكم May Allah accept your worship	إعدام قاتل فرح Execution execution	حلول عيد الفطر Eid Al-Fitr has arrived
2	عيد الفطر السعيد Happy Eid Fitr	الكويت تستحق الأفضل Kuwait deserves the best	السياحة في جورجيا Tourism in Georgia	رئيس مجلس الأمة The speaker of the national assembly	عيد الفطر المبارك Blessed Eid Fitr
3	حلول عيد الفطر Eid Al-Fitr has arrived	إعدام قاتل فرح Farah's murderer's execution	خادم الحرمين الشريفين Custodian of the tow holy mosques	عيد الفطر عيد الفطر Blessed Eid Fitr	السياحة في جورجيا Tourism in Georgia
4	تقبل الله طاعتكم May Allah accepts your worship	السياحة في جورجيا Tourism in Georgia	عيد الفطر المبارك Blessed Eid Fitr	حلول عيد الفطر Eid Al-Fitr has arrived	عيد الفطر السعيد Happy Eid Fitr
5	خادم الحرمين الشريفين Custodian of the tow holy mosques	مسجد الإمام الحسين Al-Imam Al-Hussain mosque	ابتسم انت مطير Smile you are Mutairi	عيد الفطر السعيد Happy Eid Fitr	إعدام قاتل فرح Farah's murderer's execution
6	إعدام قاتل فرح Farah's murderer's execution	رئيس مجلس الأمة The speaker of the National Assembly	مسجد الإمام الحسين Al-Imam Al-Hussain Mosque	سلمان بن عبدالعزيز Salman, son of Abdelaziz	ابتسم انت مطير Smile you are Mutairi
7	رئيس مجلس الأمة The speaker of the National Assembly	سلمان بن عبدالعزيز Salman, son of Abdelaziz	عيد الفطر السعيد Happy Eid Fitr	السياحة في جورجيا Tourism in Georgia	تقبل الله طاعتكم May Allah accept your worship
8	السياحة في جورجيا Tourism in Georgia	عيد الفطر السعيد Happy Eid Fitr	رئيس مجلس الأمة The speaker of the National Assembly	خادم الحرمين الشريفين Custodian of the tow holy mosques	مسجد الإمام الحسين Al-Imam Al-Hussain Al-mosque
9	مسجد الإمام الحسين Al-Imam Al-Hussain Mosque	عيد الفطر المبارك Blessed Eid Fitr	سلمان بن عبدالعزيز Salman, son of Abdelaziz	الكويت تستحق الأفضل Kuwait deserves the best	خادم الحرمين الشريفين Custodian of the tow holy mosques
10	سلمان بن عبدالعزيز Salman, son of Abdelaziz	تقبل الله طاعتكم May Allah accepts your worship	إعدام قاتل فرح Farah's murderer's execution	مسجد الإمام الحسين Al-Imam Al-Hussain mosque	رئيس مجلس الأمة The speaker of the National Assembly

TABLE 6. Tri-grams for the neutral posts.

#	Frequency	PMI	T-Test	Chi-Square	Likelihood
1	السياحة في جورجيا Tourism in Georgia	عيد الفطر السعيد Happy Eid Fitr	رئيس مجلس الأمة The speaker of the National Assembly	خادم الحرمين الشريفين custodian of the tow holy mosques	مسجد الإمام الحسين Al-Imam Al-Hussain mosque
2	خادم الحرمين الشريفين Custodian of the tow holy mosques	مسجد الإمام مسجد الإمام الحسين Al-Imam Al-Hussain mosque	ابتسم انت مطير Smile you are Mutairi	عيد الفطر السعيد Happy Eid Fitr	إعدام قاتل فرح Farah's murderer's execution
3	سلمان بن عبدالعزيز Salman, son of Abdelaziz	تقبل الله طاعتكم May Allah accepts your worship	إعدام قاتل فرح Farah's murderer's execution	مسجد الإمام الحسين Al-Imam Al-Hussain mosque	رئيس مجلس الأمة The speaker of the National Assembly
4	رئيس مجلس الأمة The speaker of the National Assembly	سلمان بن عبدالعزيز Salman, son of Abdelaziz	عيد الفطر السعيد happy Eid Fitr	السياحة في جورجيا Tourism in Georgia	تقبل الله طاعتكم May Allah accept your worship
5	عيد الفطر المبارك Blessed Eid Fitr	خادم الحرمين الشريفين Custodian of the tow holy mosques	تقبل الله طاعتكم may Allah accept your worship	إعدام قاتل فرح Farah's murderer's execution	حلول عيد الفطر Eid Al-Fitr has arrived
6	عيد الفطر السعيد Happy Eid Fitr	الكويت تستحق الأفضل Kuwait deserves the best	السياحة في جورجيا Tourism in Georgia	رئيس مجلس الأمة The speaker of the National Assembly	عيد الفطر المبارك Blessed Eid Fitr
7	إعدام قاتل فرح Farah's murderer's execution	رئيس مجلس الأمة The speaker of the National Assembly	مسجد الإمام الحسين Al-Imam Al-Hussain mosque	سلمان بن عبدالعزيز Salman, son of Abdelaziz	ابتسم انت مطير Smile you are Mutairi
8	مسجد الإمام الحسين Al-Imam Al-Hussain mosque	عيد الفطر المبارك Blessed Eid Fitr	سلمان بن عبدالعزيز Salman, son of Abdelaziz	الكويت تستحق الأفضل Kuwait deserves the best	خادم الحرمين الشريفين Custodian of the tow holy mosques
9	تقبل الله طاعتكم May Allah accepts your worship	السياحة في جورجيا Tourism in Georgia	عيد الفطر المبارك Blessed Eid Fitr	حلول عيد الفطر Eid Al-Fitr has arrived	عيد الفطر السعيد Happy Eid Fitr
10	حلول عيد الفطر Eid Al-Fitr has arrived	إعدام قاتل فرح Farah's murderer's execution	خادم الحرمين الشريفين Custodian of the tow holy mosques	عيد الفطر المبارك Blessed Eid Fitr	السياحة في جورجيا Tourism in Georgia

architecture of MiniLM is the same as that of the BERT model.[8]

5) CAMeLBERT [41]. Multiple BERT models are included under the CAMeLBERT, in our study, we adopt the dialectal version, which is called the ''bert-base-arabic-camelbert-da''.[9]

[8]https://huggingface.co/microsoft/Multilingual-MiniLM-L12-H384

[9]https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da

Based on the findings from previous studies, when applying an advanced deep learning language model similar to BERT models in the classification system, there will be limited effects of preprocessing Arabic text on improving the performance of the system [2], [42]. Accordingly, we used the raw text directly and no extensive preprocessing techniques were applied, only the basic data cleaning and filtering techniques that were applied during the corpus development phase were applied before using the posts to train and evaluate

TABLE 7. Tri-grams for the negative posts.

#	Frequency	PMI	T-Test	Chi-Square	Likelihood
1	مسجد الإمام الحسين Al-Imam Al-Hussain mosque	إعدام قاتل فرح Farah's murderer's execution	فرح حمزة أكبر Farah Hamza Akbar	عيد الفطر المبارك Blessed Eid Fitr	استهتار بمصير البدون Carelessness with the Bidon's future
2	فرح حمزة أكبر Farah Hamza Akbar	مسجد الإمام الحسين Al-Imam Al-Hussain Mosque	عيد الفطر المبارك Blessed Eid Fitr	استهتار بمصير البدون Carelessness with the Bidon's future	إعدام قاتل فرح Farah's murderer's execution
3	أنا الضحية القادمة I'm the next victim	استهتار بمصير البدون Carelessness with the Bidon's future	إعدام قاتل فرح Farah's murderer's execution	مسجد الإمام الحسين Al-Imam Al-Hussain Mosque	فرح حمزة أكبر Farah Hamza Akbar
4	المملكة العربية السعودية Saudi Arabia Saudi Arabia	إعدام قاتل فرح Farah's murderer's execution	مسجد الإمام الحسين Al-Imam Al-Hussain mosque	فرح حمزة أكبر Farah Hamza Akbar	عيد الفطر المبارك Blessed Eid Fitr
5	فرح حمزة أكبر Farah Hamza Akbar	عيد الفطر المبارك blessed Eid Fitr	استهتار بمصير البدون Carelessness with the Bidon's future	إعدام قاتل فرح Farah's murderer's execution	مسجد الإمام الحسين Al-Imam Al-Hussain mosque

the classification models. Moreover, all BERT models are implemented using the same parameter settings; maximum length = 128 characters, patch size = 16, epoch = 2, epsilon = 1e-8, and learning rate = 2e-5.

3) SYSTEM ARCHITECTURE AND DEVELOPMENT

All experiments are implemented using the Google Colab notebook environment and Python framework. For the baseline models, we import the Python Scikit-learn library to develop the classification system. All BERT models are imported from the HuggingFace repository and classification systems were developed in Python using the PyTorch Transformers library. We use the same dataset we developed earlier to pretrain and test models. We use the pool layer from the encoder and feed it into a simple Feed Forward Neural Network (FFNN) layer in developing the model architecture with the same parameters we defined earlier.

D. PERFORMANCE EVALUATION

In all experiments, we use macro-averaged measurements in system evaluation such as macro-average F1 score and accuracy, which help to remove any bias toward a particular class because the distribution among the three classes; positive, neutral, and negative; is not equal.

Our metrics consist of accuracy, macro-F1, Cohen Kappa score, and annotation agreement. Accuracy represents the percentage of correct predictions made by our proposed system out of the total number of predictions. Macro-F1, also called macro-averaged F1, which calculated using the arithmetic mean of all per-class F1 scores. Cohen Kappa and annotation agreement are used to measure the agreement between two raters.

We also perform hyperparameter tuning through a stratified 5-fold cross-validation approach using the train set to arrive at the most efficient hyperparameter values. Furthermore, models were evaluated using a stratified 5-fold cross-validation approach that removes bias through averaging resulting performance scores. In all experiments, we use the Scikit-Learn Python library and Google Colab notebook to implement the system evaluation measurements and metrics. In addition, an SME manually inspects the misclassified posts to provide in-depth error analysis on system performance.

IV. RESULTS

A. Q8SENTILABELER SYSTEM EVALUATION

To implement the Q8SentiLabeler system, we used a Python package for the Snorkel framework; we also created labeling functions for each ZS model.

To select the three ZS models used as labeling functions, we followed the same as in [7] and [43], from the list of ZS models published in the Hugging Face repository[10] that either support multilingual or support the Arabic language and is fine-tuned on XNLI using either XLM-R or mDeBERTa models. We excluded the models that reported poor performance and did not support the Kuwaiti dialect.

The final selected ZS models are the following:

1) joeddav/xlm-roberta-large-xnli [44]
2) MoritzLaurer/mDeBERTa-v3-base-mnli-xnli [45]
3) vicgalle/xlm-roberta-large-xnli-anli [46]

We conducted several experiments to evaluate our proposed system; the architecture of each experiment was the same as in figure 4. In this system we fixed the labels' value and changed the prompt values used in zero-shot labeling functions; we also tested the effect of combining all or partial prompts as inputs to zero-shot labeling functions in one experiment.

Tables 8 and 9 illustrate the results of the experiments; first, to evaluate the overall performance of the Q8SentiLabeler system, the results show that the average accuracy and macro-F1 are 0.90 and 0.75, respectively; Also, the average annotation agreement is 0.70, and the average Cohen Kappa score is 0.81, indicating almost perfect agreement between Q8SentiLabeler and the human labeler. We further evaluated the system by calculating the total number of generated labels; on average, the experiments were able to annotate 87% of posts with an average total of 15940 labeled posts out of 18354 unlabeled posts; the resulting labeled posts are slightly imbalanced, with more data points labeled as positive and negative than neutral, this result is the same as in [7]. The above result indicates that the Q8SentiLabeler system can effectively label Kuwaiti dialect sentiment of a large dataset of posts with good performance and quality of annotations.

The top performing experiment in all performance evaluation values is experiment 1; in this experiment, we used "The

[10]https://huggingface.co/

TABLE 8. Q8SentiLabeler System Performance Evaluation Results.

#	Prompt	Accuracy	Macro-F1	Cohen Kappa Score	Agreement
1	The sentiment of this post is	**0.92**	**0.84**	**0.87**	**0.92**
2	the tweet sentiment is { }	0.91	0.76	0.81	0.90
3	the sentiment in the tweet is { }	0.92	0.79	0.83	0.91
4	the Label of the tweet sentiment is { }	0.90	0.73	0.80	0.89
5	the sentiment label of this post is { }	0.90	0.72	0.80	0.89
6	what is the sentiment label of this post is { }	0.90	0.78	0.60	0.76
7	what is the sentiment of this post is { }	0.88	0.77	0.66	0.79
8	what is the sentiment classification of this post { }	0.87	0.71	0.56	0.73
9	tell me what is the sentiment classification label of this post { }	0.88	0.68	0.29	0.46
10	what is the sentiment classification label of this post { }	0.91	0.70	0.54	0.71
11	Experiments 1 to 10	0.89	0.75	0.81	0.89
12	Experiments 1,2,3,4,5,6,10	0.90	0.77	0.82	0.90
	Average	**0.90**	**0.75**	**0.70**	**0.81**

TABLE 9. Q8SentiLabeler System Labels Results.

#	Prompt	Positive Labels	Negative Labels	Neutral Labels	Labeled Posts
1	The sentiment of this post is { }	8264	8090	1355	17709
2	the tweet sentiment is { }	8413	8241	635	17289
3	the sentiment in the tweet is { }	8088	7922	1496	17506
4	the Label of the tweet sentiment is { }	8446	8224	810	17480
5	the sentiment label of this post is { }	8137	8799	741	17677
6	what is the sentiment label of this post is { }	7999	5817	709	14525
7	what is the sentiment of this post is { }	6965	6664	1974	15603
8	what is the sentiment classification of this post { }	7447	5847	1203	14497
9	tell me what is the sentiment classification label of this post { }	5808	2692	515	9015
10	what is the sentiment classification label of this post { }	7434	5651	346	13431
11	Experiments 1 to 10	8453	7772	2056	18281
12	Experiments 1,2,3,4,5,6,10	8418	7916	1936	18270
	Average	**7823**	**6970**	**1148**	**15940**

TABLE 10. Baseline classification models results.

	Datasets			
	Validation		Test	
	F1	Acc.	F1	Acc.
LR	0.66	0.81	0.78	0.91
SVM	**0.75**	**0.84**	**0.99**	**1.00**
M-NB	0.51	0.75	0.54	0.78
Bagging	0.67	0.76	0.98	0.99

TABLE 11. Main BERT-based classification models results.

	Datasets			
	Validation		Test	
	F1	Acc.	F1	Acc.
AraBERT	0.66	0.85	0.66	0.86
MiniLM	0.50	0.72	0.53	0.78
ARBERT	**0.72**	**0.87**	**0.75**	**0.89**
MARBERT	0.62	0.84	0.71	0.88
CAMeLBERT	0.57	0.81	0.67	0.84

sentiment of this post is " as a prompt for the three zero-shot labeling functions; the accuracy and Macro-F1 achieved were 0.92 and 0.84, the agreement with human labelers value was 0.92, and the Cohen Kappa score was 0.87; indicating that the resulting labels in experiment 1 were very high quality, the percentage of labeled posts resulted from this experiment was 96% from the unlabeled ones, with a total of 17709 out of 18354 unlabeled posts.

Besides the above results, tables 8 and 9 show that when in experiments 6 to 10 we used question format in prompts, the Cohen Kappa score values ranged from 0.29 to 0.66; in addition, the agreement with human labelers values ranged from 0.46 to 0.79; resulting in lower labeling quality than in other experiments. We also observed that the total generated labels in these experiments were lower than in other experiments as it labeled between 9015 to 15603 posts. Experiment 9 had the worst results with the Cohen Kappa score value of 0.29 and the agreement with human labelers value of 0.46, this indicates that the zero-shot models were not able to predict the labels and agree with human labelers correctly. Therefore, it can be concluded that the prompt format of using both questions and instructions does not perform well when used with the selected zero-shot models in these experiments.

Furthermore, besides testing the effect of varying prompt phrases on the performance of the Q8SentiLabeler system, we tested combining all labeling functions from experiments 1 to 10 in experiment 11. Furthermore, in experiment 12, we tested combining only the experiments that gave good accuracy results (experiments 1,2,3,4,5,6, and 10). In both experiment 11 and 12, Q8SentiLabeler system was able to label 99% of the unlabeled posts, yet when we take into consideration the other performance evaluation factors; such as accuracy, macro-F1, Cohen Kappa score, and annotation agreement the two experiments could not exceed the performance of experiment 1. However, the values were very close, indicating that with further experimentation, there is a chance to improve performance for those experiments.

TABLE 12. Misclassified Posts.

Post	True Label	Predicted Label	Analysis
سبحان الله الطفل ريان الله يرحمه وعور قلوبنا والله وفرحنا يوم طلع بس قضاء الله وحكمه الله وحطو اصوره وتكلمو عنه ناس واجد وفي مذيع طلع يبجي زين اطفال البدون ليش ماتكلمتوا عنهم الطفل براك راح يموت ولا احد قال شي والبدون الي انتحروا ماشفنا تعاطفكم شنو هالنفاق #البدون_اولويه *Glory be to God, the child Rayan, may God have mercy on him, and our hearts are broken, I swear that we were happy when he came , out but God's judgment and God's wisdom and they put pictures and many people talked about him and there was an announcer who came out crying, okay why didn't you talk about the children of the Bedoon? the child Brac will die and no one talked about him and no one said anything about Bedoon who committed suicide, we didn't see your sympathy, what is this hypocrisy #Bedoon_Priority*	Negative	Positive	This post begins with a sympathetic expression and a prayer for the child Rayan, indicating unity, however, the second part contrast that directly with a tone of anger and grief on the Bideon, which is the ultimate message of the post
العداله_لدلال_العبدالجادر قصة دلال صارت قضية المجتمع كله *Justice_for_Dalal_Al-Abduljader Dalal's story became the issue of the whole society*	Positive	Neutral	The content is a metaphor of the whole society being concerned with the victim's case, which is considered positive because it shows harmony and union
جريمة قتل فرح خلت كل محقق و وكيل نيابه وقاضي يعيد حساباته بقرارات التحقيق اللي كان ياخذها بالقضايا وكانها صفعة عشان تصحينا وتخلينا نستخدم السقف الاعلى من صلاحيات التحقيق عزاء_النساء #فرح_حمزة أكبر *Farah's murder made every investigator, prosecutor, and judge recalculate his accounts of the investigation decisions he used to take in the cases, as if it was a slap to wake us up and let us use the higher ceiling of investigation powers. The condolences of use the higher ceiling of investigation powers. The condolences of the women, #Farah_Hamza_Akber*	Neutral	Negative	The statement is an opinion that is expressed in a neutral tone without negativity. The word صفعة that means (slap) can indicate anger in other contexts, but it is only used for emphasis in this text. The mention of (murder /جريمة قتل) could be related to the negative classification as well
عبدالكريم الكندري: على وزير الداخلية إجراء تحقيق في حادثة التعدي على رجل دين #وزارة_الداخلية #الكويت #السعودية عيد_الفطر #زكاة_الفط *Abdul Karim Al-Kandari: The Minister of Interior should conduct an investigation into the incident of assaulting a cleric Ministry of Interior Kuwait Saudi Arabia #Eid_al-Fitr #Zakat_al-Fitr*	Neutral	Negative	In the statement, the opinion is given in a neutral tone, with no emotions that convey negativity. It may have been classified as negative due to wordings (تحقيق\حادثة\تعدي) which reflect an accident

Finally, evaluating classes distribution of the labeled dataset, most experiments generated a slightly imbalanced dataset, with more data points labeled as positive and negative than neutral.

B. CLASSIFICATION SYSTEMS PERFORMANCE

The results of the baseline models are reported in Table 10. As can be observed, in both the validation set and test set, accuracy, and macro-averaged F1 scores are the highest for the SVM classifier among the baseline models. It demonstrates perfect performance for the test set with 0.99 and 1.00 for the macro-averaged F1 and accuracy scores respectively. These perfect performance scores imply a possibility of over-fitting. After further analysis of the SVM results, we found only three misclassified positive posts that are labeled as negative, five negative posts are misclassified as positive, only one neutral post is misclassified as positive, and four neutral posts are misclassified as negative. A similar finding is also applied to the bagging model.

Table 11 records the results of the main BERT-based classification models. Even though AraBERT includes in its pre-training corpus posts and emoji, similar to our dataset, and MARBERT is constructed using a large posts dataset as well, they both were not performing as well as the ARBERT model. The ARBERT model, pre-trained using the MSA corpus, reports 0.75 and 0.89 for the macro-averaged F1 and accuracy scores respectively on the test set.

C. ERROR ANALYSIS

This section narrows our scope to focus on the best-performance BERT model, the ARBERT model. We retrieve some samples from the misclassified posts and investigate their content manually. Table 12 lists some samples from the misclassified posts including some analysis of their content, which might mislead the classification model and create some confusion.

Although the BERT model is a contextual aware model that analyzes both the content and context of the posts, the predicted labels from the classification model depend heavily on the wording of the posts. Based on our manual investigation, texts with positive vocabulary were considered positive, and those with harsh and aggressive words were classified as negative, rather than the hidden meaning and emotions behind it. This is related to abstract concepts that are difficult to detect unless the reader understands social rules and can pick up on hints of irony, sarcasm, humor, and other similar perceptions in the Kuwaiti dialect.

V. FUTURE WORK

Our next step is to apply our proposed Q8SentiLabeler System to various Arabian Gulf dialects as that can help to examine its validity on other dialects that share similar linguistic features with the Kuwaiti dialect. We also would like to improve our proposed Q8SentiLabeler System by advancing its structure to include more ZS models and knowledge sources that can provide valuable input to the model.

As more online users are expressing theirself in their native dialect, it is becoming very important to apply especialized Kuwaiti NLP applications in Kuwait. Our proposed system supports multiple NLP applications, such as online forums and platforms moderation and monitoring, track customer sentiment and behavior online, social problem analysis, and product analysis.

VI. CONCLUSION

This study proposed the "Q8SentiLabeler", a weak supervised sentiment analysis system for the Kuwaiti dialect. The main goal of the "Q8SentiLabeler" system is to save time, cost, and effort required by hiring human annotators. Accordingly, we were able to develop the first large corpus for sentiment analysis of the Kuwaiti dialect. Furthermore, we evaluate our dataset by applying several traditional machine-learning classifiers and advanced deep-learning language model classifiers. The results demonstrate the positive impacts and potential of applying our proposed framework to achieve high-performance accuracy. In general, our proposed framework supports the advancement of research in NLP and the robust creation of linguistic resources by diminishing the need for human annotators.

REFERENCES

[1] N. Y. Habash, *Introduction to Arabic Natural Language Processing* (Synthesis Lectures on Human Language Technologies), vol. 3, 1st ed. San Rafael, CA, USA: Morgan & Claypool, 2010, doi: 10.2200/S00277ED1V01Y201008HLT010.

[2] F. Husain and O. Uzuner, "Investigating the effect of preprocessing Arabic text on offensive language and hate speech detection," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 4, pp. 1–20, Jan. 2022, doi: 10.1145/3501398.

[3] F. Husain and O. Uzuner, "A survey of offensive language detection for the Arabic language," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–44, Mar. 2021, doi: 10.1145/3421504.

[4] S. Al-Qenaie, "Kuwaiti Arabic: A socio-phonological perspective," Ph.D. dissertation, School Mod. Lang. Cultures, Durham Univ., Durham, U.K., 2011.

[5] N. A. Hayat and Y. B. AlBader, "The McChicken phenomenon: How has English become a prevalent language among Kuwaiti youths?" *World J. English Lang.*, vol. 12, no. 6, p. 59, Jul. 2022.

[6] J. B. Salamah and A. Elkhlifi, "Microblogging opinion mining approach for Kuwaiti dialect," in *Proc. Int. Conf. Comput. Technol. Inf. Manage. (ICCTIM)*, Dubai, Dubai, 2014, p. 388.

[7] F. Husain, H. Al-Ostad, and H. Omar, "A weak supervised transfer learning approach for sentiment analysis to the Kuwaiti dialect," in *Proc. The 7th Arabic Natural Lang. Process. Workshop (WANLP)*, Abu Dhabi, UAE, 2022, pp. 161–173. [Online]. Available: https://drive.google.com/file/d/11-ae50ZgvyQsQYZ3ZS0N2yj_xAG96WEy/view

[8] H. Aldihan, R. Gaizauskas, and S. Fitzmaurice, "A pilot study on the collection and computational analysis of linguistic differences amongst men and women in a Kuwaiti Arabic WhatsApp dataset," in *Proc. 7th Arabic Natural Lang. Process. Workshop (WANLP)*, Abu Dhabi, UAE, 2022, pp. 372–380.

[9] I. Abu Farha and W. Magdy, "The effect of Arabic dialect familiarity on data annotation," in *Proc. The 7th Arabic Natural Lang. Process. Workshop (WANLP)*, Abu Dhabi, United Arab Emirates, 2022, pp. 399–408.

[10] D. Cacciarelli and M. Kulahci, "Active learning for data streams: A survey," *Mach. Learn.*, vol. 113, no. 1, pp. 185–239, Nov. 2023, doi: 10.1007/s10994-023-06454-2.

[11] R. Jiao, Y. Zhang, L. Ding, B. Xue, J. Zhang, R. Cai, and C. Jin, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," *Comput. Biol. Med.*, vol. 169, Feb. 2024, Art. no. 107840. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523013057

[12] W. Tok, A. Bahree, and S. Filipi, *Practical Weak Supervision: Doing More With Less Data*. O'Reilly Media, 2021.

[13] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *Proc. VLDB Endowment*, vol. 11, no. 3, pp. 269–282, Nov. 2017.

[14] M. Chen, "Analysis on transfer learning models and applications in natural language processing," *Highlights Sci., Eng. Technol.*, vol. 16, pp. 446–452, Nov. 2022.

[15] L. Tunstall, L. von Werra, and T. Wolf, *Natural Language Processing With transformers*. O'Reilly Media, 2022.

[16] S. Yildirim and M. Asgari-Chenaghlu, *Mastering Transformers: Build State-of-the-art Models From Scratch With Advanced Natural Language Processing Techniques*. Birmingham, U.K.: Packt Publishing, 2021. [Online]. Available: https://books.google.com.kw/books?id=A9Y6EAAAQBAJ

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[18] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," 2021, *arXiv:2111.09543*.

[19] F. Husain, "Arabic offensive language detection in social media," Ph.D. dissertation, Dept. Inf. Sci. Technol., George Mason Univ., Fairfax, VA, USA, 2021.

[20] E. Tsakiri, "Semi-supervised sentiment analysis for sentence classification," Master's thesis, Dept. Linguistics Philol., Uppsala Univ., Uppsala, Sweden, 2022.

[21] G. Tripathy and A. Sharaff, "A hybridized Ga-based feature selection for text sentiment analysis," in *Encyclopedia of Data Science and Machine Learning*. Hershey, PA, USA: IGI Global, 2023, pp. 1858–1870.

[22] K. Sandar Kyaw, P. Tepsongkroh, C. Thongkamkaew, and F. Sasha, "Business intelligent framework using sentiment analysis for smart digital marketing in the e-commerce era," *Asia Social Issues*, vol. 16, no. 3, Jan. 2023, Art. no. e252965.

[23] R. Mahdalina and A. Wicaksono, "E-commerce customer satisfaction analysis on microblogs," *Syntax Idea*, vol. 5, no. 1, p. 11, Jan. 2023. [Online]. Available: https://www.jurnal.syntax-idea.co.id/index.php/syntax-idea/article/view/2091

[24] F. Cardoso Durier da Silva, A. C. Bicharra Garcia, and S. W. M. Siqueira, "Sentiment gradient, an enhancement to the truth, lies and sarcasm detection," in *Advances in Artificial Intelligence—IBERAMIA 2022*, A. C. B. Garcia, M. Ferro, and J. C. Rodríguez Ribón, Eds. Cham, Switzerland: Springer, 2022, pp. 107–118.

[25] R. Feldman, B. Rosenfeld, R. Bar-Haim, and M. Fresko, "The stock sonar—Sentiment analysis of stocks based on a hybrid approach," in *Proc. AAAI Conf. Artif. Intell.*, Aug. 2011, vol. 25, no. 2, pp. 1642–1647. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/18854

[26] J. Shah, M. Shah, H. Gala, J. Hirpara, and P. A. Gawade, "YouTube assistant for quick navigation and content moderation," in *Proc. IEEE 4th Int. Conf. Adv. Electron., Comput. Commun. (ICAECC)*, May 2022, pp. 1–6.

[27] T. M. Omran, B. T. Sharef, C. Grosan, and Y. Li, "Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach," *Data Knowl. Eng.*, vol. 143, Jan. 2023, Art. no. 102106. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X22000970

[28] A. A. Al Shamsi and S. Abdallah, "Sentiment analysis of emirati dialect," *Big Data Cognit. Comput.*, vol. 6, no. 2, p. 57, May 2022. [Online]. Available: https://www.mdpi.com/2504-2289/6/2/57

[29] S. Rizkallah, A. Atiya, H. ElDin Mahgoub, and M. Heragy, "Dialect versus MSA sentiment analysis," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.* Cham, Switzerland: Springer, 1007, pp. 605–613.

[30] R. M. Alahmary, H. Z. Al-Dossari, and A. Z. Emam, "Sentiment analysis of Saudi dialect using deep learning techniques," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Jan. 2019, pp. 1–6.

[31] M. Alruily and O. R. Shahin, "Sentiment analysis of Twitter data for Saudi universities," *Int. J. Mach. Learn. Comput.*, vol. 10, no. 1, pp. 18–24, Jan. 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216315702
[32] B. A. Alharbi, M. A. Mezher, and A. M. Barakeh, "Tourist reviews sentiment classification using deep learning techniques: A case study in Saudi Arabia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 717–726, 2022, doi: 10.14569/ijacsa.2022.0130685.
[33] C. Manning and H. Schütze, *Foundations of Statistical Natural Language*. Cambridge, MA, USA: MIT Press, 1999.
[34] F. Husain and O. Uzuner, "SalamREPO: An Arabic offensive language knowledge repository," in *Proc. 4th Int. Conf. Comput. Appl. Inf. Secur.*, 2021, pp. 1–6.
[35] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of Hindi–English code-mixed social media text for hate speech detection," in *Proc. 2nd Workshop Comput. Modeling People's Opinions, Personality, Emotions Social Media*, New Orleans, LA, USA, 2018, pp. 36–41. [Online]. Available: https://aclanthology.org/W18-1105
[36] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, 1145, pp. 145–153, doi: 10.1145/2872427.2883062.
[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, p. 2.
[38] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. Workshop Lang. Resour. Eval. Conf.*, May 2020, p. 9.
[39] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 7088–7105.
[40] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5776–5788.
[41] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Kyiv, Ukraine, Apr. 2021.
[42] F. Husain, "OSACT4 shared task on offensive language detection: Intensive preprocessing-based approach," in *Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools, Shared Task Offensive Lang. Detection*. Marseille, France, May 2020, pp. 53–60. [Online]. Available: https://aclanthology.org/2020.osact-1.8
[43] H. Alostad, S. Dawiek, and H. Davulcu, "Q8 VaxStance: Dataset labeling system for stance detection towards vaccines in Kuwaiti dialect," *Big Data Cognit. Comput.*, vol. 7, no. 3, p. 151, Sep. 2023. [Online]. Available: https://www.mdpi.com/2504-2289/7/3/151
[44] J. Davison. *XLM-RoBERTa-Large-XNLI*. Accessed: Feb. 16, 2024. [Online]. Available: https://huggingface.co/joeddav/xlm-roberta-large-xnli
[45] M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers, "Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI," *Political Anal.*, vol. 32, no. 1, pp. 84–100, Jan. 2024.
[46] V. Gallego. *XLM-RoBERTa-Large-XNLI-ANLI*. Accessed: Feb. 16, 2024. [Online]. Available: https://huggingface.co/vicgalle/xlm-roberta-large-xnli-anli

FATEMAH HUSAIN is an Assistant Professor with the Information Science Department, Kuwait University. Her research focuses on applications and technological solutions to foster social justice. In particular, she develops multiple natural language processing systems to detect offensive language and hate speech from social media content. Her research also involves investigating the effect of social media on users' health.

HANA ALOSTAD received the Ph.D. degree in computer science from Arizona State University. She is a Lecturer with the Gulf University for Science and Technology (GUST). Her Ph.D. research area focuses on NLP, social network analysis, and its applications in algorithmic trading. Currently, she is more focusing on her research projects on Arabic NLP for the Kuwaiti dialect. Besides her academic career, she has more than 15 years of practical experience in software engineering and project management, particularly in projects related to governmental financial and accounting systems.

HALIMA OMAR is currently pursuing the degree majoring in communication disorders with Kuwait University. She knows first-hand the frustration of the shortage of Arabic resources, which formed the basis of her passion for expanding and employing her knowledge to contribute to the enhancements in this field. Among her duties as a professor's assistant with the university, she tutors fellow students in linguistics and language development. She strongly believes in the value of the intersection of various fields in creating advanced and impactful research. She has received several honors for her work and academic achievement.