A Weak Supervised Transfer Learning Approach for Sentiment Analysis to the Kuwaiti Dialect

Fatemah Husain
Kuwait University
College of Life Sciences
Information Science Department
f.husain@ku.edu.kw

Hana Al-Ostad
Gulf University for Science & Technology
College of Arts & Sciences
Computer Science Department
alostad.h@gust.edu.kw

Halima Omar
Kuwait University
College of Life Sciences
Communication Disorders Science Department
halima.omar@cls.ku.edu.kw

Abstract

Developing a system for sentiment analysis is very challenging for the Arabic language due to the limitations in the available Arabic datasets. Many Arabic dialects are still not studied by researchers in Arabic sentiment analysis due to the complexity of annotators' recruitment process during dataset creation. This paper covers the research gap in sentiment analysis for the Kuwaiti dialect by proposing a weak supervised approach to develop a large labeled dataset. Our dataset consists of over 16.6k tweets with 7,905 negatives, 7,902 positives, and 860 neutrals that spans several themes and time frames to remove any bias that might affect its content. The annotation agreement between our proposed system's labels and human-annotated labels reports 93% for the pairwise percent agreement and 0.87 for Cohen's kappa coefficient. Furthermore, we evaluate our dataset using multiple traditional machine learning classifiers and advanced deep learning language models to test its performance. The results report 89% accuracy when applied to the testing dataset using the ARBERT model.

1 Introduction

Datasets are the foundation of the most significant innovation in the field of Natural Language Processing (NLP). The development of NLP algorithms and tools is dependent on the availability and quality of the datasets that serve their goals. While there are plenty of English language datasets, for some other natural languages, there are still minimal resources, such as the Arabic language (Husain and Uzuner, 2021). The Arabic language is considered among the low-resource languages for NLP, however, the number of people who speaks Arabic exceeds 353.6 million [1].

The Arabic language has multiple forms. The Classical Arabic Language (CAL) is the oldest form of Arabic and is often used in Islamic manuscripts (e.g., the Quran) (Habash, 2010; Husain and Uzuner, 2022). Modern Standard Arabic (MSA) is the official language for Arabic countries and it is used in official media resources, writing books, etc(Habash, 2010; Husain and Uzuner, 2021). The last and most dominant form of Arabic is the Arabic dialects, which are the native language form of daily communication. The Arabic dialects differ based on geographical and social classes(Habash, 2010). Moreover, Arabic dialects are often used in online user-generated content such as on Twitter, Facebook, and Instagram. This variation among Arabic dialects makes it very challenging to develop tools that can process Arabic social media content accurately.

In this study, we develop a dataset based on an innovative method to reduce the number of human annotators and propose a text classification model for sentiment analysis specifically for the Kuwaiti dialect. The Kuwaiti dialect has not been comprehensively covered and studied in previous computational linguistic research. According to our knowledge, only(Salamah and Elkhlifi, 2014) investigates some linguistic tools for the Kuwaiti dialect to develop an approach for unsupervised sentiment analysis, however, their dataset is not publicly available for researchers. This gap in research inspires us to further study the Kuwaiti dialect and

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

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 161 - 173

to create linguistic resources to support research in this area. This initial step in studying the Kuwaiti dialect could also support the study of other underrepresented Arabian Gulf dialects that might share some vocabularies with the Kuwaiti dialect, for example, it can help researchers in Bahraini or Qatari dialects.

The key contributions of this study are threefold:

1. Introducing the first public Kuwaiti dataset for sentiment analysis with over 16.6K tweets covering various topics.

2. Implementing a unique data labeling system inspired by (Smith et al., 2022) for the **language model in a loop by incorporating prompting into weak supervision**, which combines the benefits of using weak supervised learning and zero-shot pre-trained transfer learning models.

3. Comparing the performance of multiple classical machine learning classifiers and several BERT models for sentiment analysis covering the Kuwaiti dialect.

This paper starts with some background information after the introduction that covers the Kuwaiti dialect, sentiment analysis resources, the latest approaches and software frameworks in labeling large datasets, the weak supervised techniques, and the zero-shot models applied in the experiments. The methodology is discussed in detail in the third section, including dataset construction, dataset labeling, classification model, and performance evaluation. In the third section, we present the results, error analysis, and discuss them thoroughly. The paper concludes with a conclusion and proposes directions for future works. The paper also includes an ethics statement at the end and appenedices.

2 Background

2.1 The State of Kuwait and the Kuwaiti Dialect

The state of Kuwait is a small country with a total area of 17,820 square kilometers located in the northwestern corner of the Persian Gulf (i.e. Arabian Gulf). Geographically, Kuwait was divided into four main areas; Sharq (East), Qibla (West), Hay al−Wasat (Middle Neighbourhood), and al−Mirqab (South)(Al-Qenaie et al., 2011). 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(Al-Qenaie et al., 2011). Furthermore, Kuwait was a protectorate of the British Empire for 62 years, which also create an effect on the Kuwaiti dialect(Hayat and AlBader, 2022). This complex structure of the Kuwaiti dialect makes it very difficult to create a linguistic system that can automatically process Kuwaiti text accurately.

2.2 Sentiment Analysis Datasets

The available research in sentiment analysis for the Kuwaiti dialect is very limited. Salamah and Elkhlifi(Salamah and Elkhlifi, 2014) create a dataset of 340,000 tweets related to the interrogation of ministers by the National Assembly of Kuwait. Other Arabian gulf dialects have also been recently targeted to develop sentiment analysis datasets. A parallel balanced dataset of English, MSA, and Bahrani dialect consisting of 5,000 product reviews and a dataset of 500 movie comments in Bahraini dialect were created for a sentiment analysis system(Omran et al., 2022). In (A. Al Shamsi and Abdallah, 2022), the authors introduced the first Emirati sentiment analysis dataset, which consists of 70,000 Instagram comments. Multiple sentiment analysis resources were developed for the Saudi dialect, such as: (1) (Rizkallah et al., 2018) develop 2010 tweets dataset for sentiment analysis; (2) (Alahmary et al., 2019) collect 32,063 Saudi tweets; (3) (Alruily and Shahin, 2020) construct a dataset of 11,764 tweets about Saudi universities; (4) in (Alharbi et al., 2022), the authors create a dataset of 22,433 reviews of tourist places.

2.3 Labeling Large Training Dataset

Data labeling is one of the most challenging tasks in creating datasets for text classification. The following points summarize the main challenges in NLP related to data labeling:

- Advanced deep learning and transfer learning algorithms require very large size labeled datasets.

- Subject Matter Experts (SMEs) have limitted time, thus its difficult to obtain labels for a large dataset from SMEs.

- In the case of crowd-sourcing, the labeling task will be very costly and raise some quality

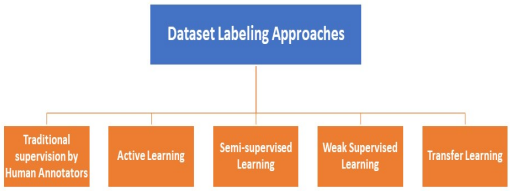

Figure 1: Summary of data labeling approaches

issues (e.g., proficiency in the subject, personal bias, background knowledge effects, and agreement among annotators).

- Privacy might be required in some projects, which might impact the annotation process and the recruitment of annotators.

Knowing the complexity behind the labeling process, researchers proposed many solutions to label data without human annotators. Fig.1 illustrates a summary of different approaches to labeling/annotating data, including both with and without help from SMEs.

Active learning is one of the advances in the traditional labeling by SMEs for supervised learning. It attempts to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle (e.g., a human annotator). In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data(Settles, 2009).

The second approach of data labeling is **semi-supervised learning**, based on (Ratner et al.; Engelen and Hoos, 2020) in this approach a small dataset is labeled using an unsupervised algorithm, then the small dataset is used to label a much larger unlabeled dataset.

The third approach is **transfer learning**, based on (Pan and Yang, 2010); this approach aims to extract the knowledge from one or more source tasks (model pre-trained on a different dataset) and apply the knowledge to a target task (to label the dataset).

The above three approaches reduce the need for SMEs to annotate additional training datasets. However, using these approaches will not avoid the need to label some data; this will not be the case when using **weak supervised learning** or **Zero-Shot (ZS) learning**, where the first approach avoid human labeling by using labeling functions created with the help of the SMEs who provides supervision at a higher level than case-by-case labeling, and the ZS learning make use of pre-trained model to label the dataset without any additional fine-tuning on the new corpus (Tunstall et al., 2022).

2.3.1 Weak supervised learning

Weak supervised learning is defined by (Tok et al., 2021) as a collection of techniques in machine learning in which models are trained using sources of information that are easier to provide than hand-labeled data, where this information is incomplete, inexact, or otherwise less accurate.

The noisy, weak labels are combined using a generative model trained based on the accuracies of the labeling functions; the accuracies are derived from agreement and disagreement of the labeling functions and used to form the training data.

Weak supervision has received much attention in recent years, and several open-source software frameworks for weak supervision have been released to be used by data scientists in building real-world systems. Example software frameworks include Snorkel(Ratner et al., 2017), SwellShark(Biomedical NER)(Fries et al., 2017), and FlyingSquid(Fu et al., 2020).

Stanford researchers, found that when they compared to the productivity of teaching the SMEs Snorkle weak supervised framework, versus spending the equivalent time just hand-labeling data, the team was able to build models not only 2.8x faster but also with 45.5% better predictive performance on average(Ratner et al., 2017). Also, they found that using Snorkel leads to an average of 132% performance improvement over baseline techniques(Ratner et al., 2017).

Another research on weak supervised learning by MIT researchers found that the combination of a few "strong" labels and a larger "weak" label dataset resulted in a model that learned well and trained at a faster rate(Robinson et al., 2020).

2.3.2 Snorkel Open Source Weak Supervision Framework

Snorkel framework(Ratner et al., 2017) is a project proposed by researchers at Stanford AI Lab started in the year 2015. It is the oldest among the weak supervised learning software frameworks. Snorkel team published over 60+ peer-reviewed publications(AI). Besides the open-source library, the Snorkel research team built a commercial ver-

sion called **Snorkel Flow**[2] by incorporating years of experience from applying weak supervision to real-world machine learning problems.

The following describes the steps of the Snorkel system:

1. The SME users write Labeling Functions (LFs) that express weak supervision sources like distant supervision, patterns, and heuristics.

2. Snorkel applies the LF on unlabeled data and learns a generative model to combine the LFs' outputs into probabilistic labels.

3. Snorkel uses these labels to train a discriminative classification model, such as a deep neural network.

2.3.3 Zero-Shot (ZS) Learning

Based on (Tunstall et al., 2022) ZS classification is suitable in a setting where no labeled data is provided. Using Natural Language Inference (NLI) the ZS model can predict the class of the unlabeled sample, even if the model was not trained on those classes. The ZS models leverage the semantic similarity between labels and the text context(Yildirim and Asgari-Chenaghlu, 2021). In this type of experiment setup, the text is treated as the premise, and the hypothesis is formed as "this example is about {label}". In addition, a set of expected labels is fed to the promise, and the entailment score tells if the promise is about that topic/label or not.

A good candidate to perform ZS classification on languages other than English is XLM-RoBERTA (XLM-R) model. It was trained on one hundred languages, including Arabic and many other low-resource languages. Based on the findings from (Conneau et al., 2020), applying the XLM-R model on the cross-lingual Natural Language Inference (XNLI) task, significantly outperforms multilingual BERT (mBERT) by +13.8% average accuracy. Moreover, it also performs exceptionally well on low-resource languages, improving 11.8% in XNLI accuracy for Swahili and 9.2% for Urdu over the previous XLM model.

Another State-Of-The-Art (SOTA) model in XNLI task is Multilingual mDeBERTa. As of December 2021, mDeBERTa-base is the best performing multilingual base-sized transformer model, it achieved a 79.8% ZS cross-lingual accuracy on XNLI and a 3.6% improvement over XLM-R Base (He et al., 2021).

2.4 Language Models in a Loop

In (Smith et al., 2022), the researchers proposed a framework incorporating ZS model prompting into programmatic weak supervision. The following is a detailed explanation of the steps:

1. The SMEs express their domain knowledge via prompts combined with unlabeled examples and given to a pre-trained ZS language model.

2. The ZS model's responses are interpreted with label maps to produce votes on the true label.

3. These votes are denoised with a label model, and the resulting estimated labels are used to train an end model.

4. The SMEs can refine their prompts throughout the process by inspecting unlabeled examples and evaluating with a small labeled development set.

Based on the findings from (Smith et al., 2022), using this approach which combines ZS models with weak supervised learning, can significantly improve performance over using the ZS model alone, with an average of 19.5% reduction in errors. They also found that this approach produces classifiers with comparable or superior accuracy to those trained from hand-engineered rules.

3 Methodology

3.1 Dataset

3.1.1 Dataset Extraction, Collection, and Filtering

The process used in collecting data spans over one year to ensure the diversity of data content, and to remove any bias or impacts that might be caused by social factors within the Kuwaiti society. We select four controversial events that happen in different time frames in Kuwait. These events create debatable and stressful content on the online Arabic Twitter-sphere. The followings are a short description of each event and the hashtags used to extract its tweets:

- Farah Akbar. These tweets were collected during April 2021. Farah Akbar is a Kuwaiti woman who was brutally murdered. Her

[2]https://snorkel.ai/

killer had threatened and harassed her after she rejected his marriage proposal. The hashtags used to extract these tweets related to Farah's event are:#عزاء_النساء and #السالم_صباح_قتل_جريمة.

- Dalal Al-Abd Al-Jader. These tweets were collected during 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 tweets is #العبدالجادر_لدلال_العدالة.

- Bideon. Bidoon or bedun refers to a stateless Arab minority in Kuwait. They do not have nationalities and are not allowed to obtain most official documents, which causes difficulties in finding employment, accessing healthcare, and education. We select tweets that were posted during February 2022 about the Bidoon because it coincides with the Moroccan child Rayan incident which received the attention of an overwhelming number of online users including Kuwaitis. This reaction from Kuwaitis toward Rayan incident increased the anger of people from the Bidoon community in Kuwait, which led them to go out to the streets and protest for their citizenship and other civil rights. The hashtags used to extract these tweets are #البدون_اولويه #البدون_عبدالعزيز#البدون_البدون_الطفل.

- Sheick Al-Hazem. These tweets were collected during April 2022. Sheikh Al-Hazem is a Kuwaiti Shia clergy who was assaulted while in the mosque by three government officials who try to confiscate money collected from people for Zakat (i.e. donation). The hashtag used to extract these tweets is #محشوم_الشيخ_مهدي_الهزيم.

3.1.2 Dataset Labeling

Tweets are categorized according to the feeling in which they are present, either to be positive; such as happiness, fun, and pride, or to be negative; such as sadness and contempt, or to be neutral in the sense that there is no expression of feelings. The followings are samples from the dataset from each label:

- Positive: وجود كل انسان اليوم بساحة الارادة حسسني بالامان

  "People's presence at the Will Square today makes me feel safe".

- Neutral: حملة مناهضة العنف ضد المرأة

  "The campaign against women's violence".

- Negative: لاشئ يؤلم اكثر من خيبة امل تاتيك من شخص ظننت انه لن يؤذيك ابدا

  "Nothing hurts more than disappointment comes from someone you thought would never hurt you".

Snorkel and Language Models in a Loop for Dataset Labeling:

We used Snorkel open-sourced software framework(Ratner et al., 2017, 2016) because the available alternative frameworks are not supporting our goal. For example, SwellShark is used for Biomedical NER, Skweak is tightly integrated with SpaCy which does not support Arabic, and FlyingSquid has limited documentation with a focus on video classification.

Fig.2 illustrates the steps we followed to label the training dataset. Our proposed labeling system differs from (Smith et al., 2022) system as for the LFs, we used several ZS pre-trained models and one promote instead of using one ZS model and changing the promote as in (Smith et al., 2022).

To select the ZS pre-trained models used in our experiments, firstly, we searched for the top ZS pre-trained models published in the Hugging Face repository[3]. The selection criteria were based on the list of top downloaded ZS models that either support multilingual or support the Arabic language and is fine-tuned on XNLI using either XLM-R or mDeBERTa models. We applied this selection criteria because any ZS model fine-tuned on one of those two models is expected to give good result with low-resource languages such as Arabic dialects as previous studied demonstrated(Conneau et al., 2020; He et al., 2021). Next, we tested the previously selected models using part of our dataset. We excluded the models that reported poor performance and did not support the Kuwaiti dialect.

After extensive experimenting, the final selected ZS models are the following:

1. joeddav/xlm-roberta-large-xnli(Davison) [4]

[3]https://huggingface.co/

[4]https://huggingface.co/joeddav/xlm-roberta-large-xnli

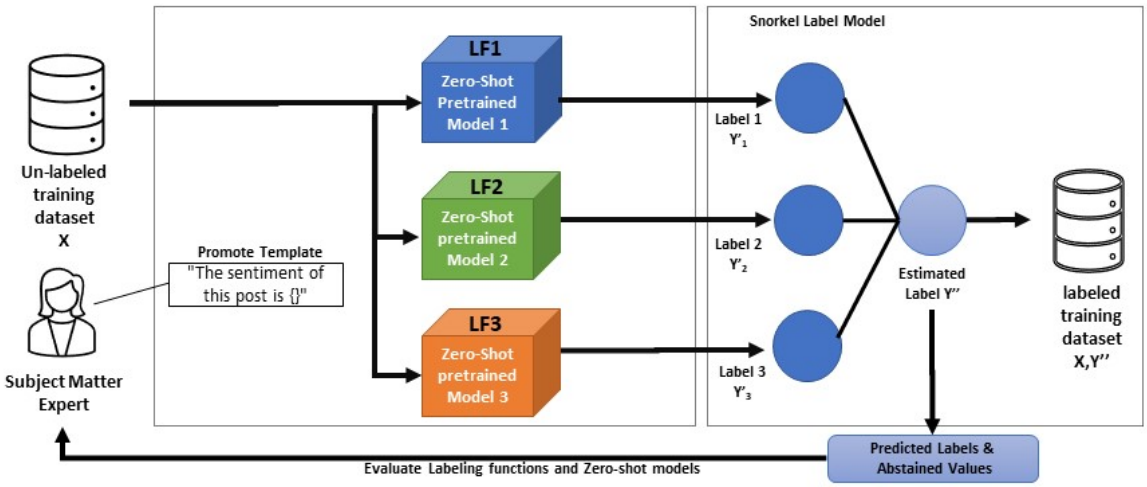

Figure 2: Snorkel weak supervised learning steps

2. MoritzLaurer/mDeBERTa-v3-base-mnli-xnli(Laurer et al., 2022) [5]

3. vicgalle/xlm-roberta-large-xnli-anli (Davison) [6]

Using the selected ZS models, we created three LFs, the LF either returns a sentiment label (positive, negative, neutral) or returns the "ABSTAIN" value in case the labeling function could not label the text. We also set the promote hypothesis template to "The sentiment of this post is {}".

Next, we applied the LF to the unlabeled training dataset. We iterated on this process several times. In each iteration, we checked the abstained tweets and samples of the predicted tweets to evaluate and refine the sentiment labels keywords and the ZS language models.

Then, we tested the performance of the Snorkel probabilistic labeling model based on the exact steps illustrated in Fig.2, but we applied it to the gold-labeled testing dataset. Finally, we retrieved the resulting labeled training dataset by removing the abstained tweets and keeping only the labeled tweets.

Gold-Labeled Dataset: In addition to Snorkel's labeled dataset, we hire 7 annotators between the age of 17 and 24 years who are Kuwaiti and proficient in the Kuwaiti dialect among other Arabic dialects to manually label a set of 2,100 tweets (300 tweets per annotator). A detailed labeling instruction including definitions and samples from each label along with a background survey and a pilot study were used to help the annotators to provide accurate labels. The pilot study consist of 15 tweets; 5 were labeled as samples from different labels and 10 were used to test the annotators. Annotators who accurately labeled the testing 10 tweets were presented with 300 tweets to label as part of the gold-labeled dataset.

We further check the human-labeled tweets for accuracy by reviewing them with an expert annotator and excluding all inexact tweets. The final version of the gold-labeled dataset consists of 1,534 tweets. This set of tweets was used to further examine our approach to data labeling using weak supervision techniques.

3.1.3 Dataset Cleaning and Preprocessing

We removed duplicated tweets, retweet keyword "RT", and user mentions. Previous studies highlighted the limited effects of preprocessing Arabic tweets when used with advanced classification models such as BERT-model(Husain and Uzuner, 2022; Husain, 2020). Thus, only hashtags were removed before applying feature extractions and using the text for the classification models.

The size of the resulting labeled dataset from our proposed labeling system, and after removing the abstained tweets is a total of 16,667 tweets; 7,905 negative, 7,902 positive, and 860 neutral. The resulting labeled dataset is nearly balanced on tweet counts between negative and positive labels, but not on the neutral labels. At this stage, the labeled dataset is ready for the next step to be used in baseline models and to fine-tune Arabic language

[5]https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

[6]https://huggingface.co/vicgalle/xlm-roberta-large-xnli-anli

models.

3.2 Classification Models

We randomly split the dataset into three parts; the train set with 60% of the total number of tweets, the validation set with 20%, and the test set is 20%. All sets have equal proportions of label distributions, Fig.3 shows the distribution of each set. Firstly, we train the classification models using the train set and evaluate them using the validation set, then we combine the validation set with the train set and train the classification models and evaluate them using the test set. As described in the following sections, multiple classifiers were applied to evaluate the dataset.

3.2.1 Baseline Models

We develop four baseline classification models; Logistic Regression (LR), Support Vector Machine (SVM), Multinomial Naive Bayes (M-NB), and Bagging with a 2-5 characters-based TF-IDF vectorizer. Previous studies emphasize the importance of applying a character-based feature when the dataset is extracted from user-generated content such as Twitter because character-based features are language-independent features that perform well with misspelling errors or obfuscating words, as is the case on most Twitter content(Bohra et al., 2018; Nobata et al., 2016). The feature and models were implemented using Python scikit-learn library.

3.2.2 BERT Models

The main classification models which we used in developing and evaluating the sentiment analysis system are sharing the same Bidirectional Encoder Representations from Transformers (BERT) architecture, however, they vary in the parameters and data used in creating them. The BERT model applies pre-trained language representations to downstream tasks through a fine-tuning approach. This approach is also called transfer learning, in which the pre-trained language representations are developed using a neural network model on a known task, and then fine-tuning is performed to use the same model for a new purpose-specific task such as sentiment analysis(Devlin et al., 2018). The following four BERT models are applied in our experiments:

- AraBERT Model(Antoun et al.). 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 tweets. This version of AraBERT includes emoji in its vocabulary[7].

- ARBERT(Abdul-Mageed et al., 2021). It uses the same network architecture of the BERT base model with a large MSA dataset that has been collected from 6 various sources[8].

- MARBERT(Abdul-Mageed et al., 2021). 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 tweets. It has the same architecture as ARBERT, but without the Next Sentence Prediction (NSP) objective as tweets are concise and short.

- Microsoft Multilingual Model (MiniLM)(Wang et al., 2020). 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 architecture of MiniLM is the same as that of the BERT model[9].

All BERT models used in this study were from the Hugging Face repository and the experiment was developed in Python using the PyTorch- Transformers library. The models were used with the same parameters settings; maximum length = 128 characters, patch size = 16, epoch = 2, epsilon = 1e-8, and learning rate = 2e-5. We did not use feature engineering because fine-tuning and deep learning do not need feature engineering, instead, we use the pool layer from the encoder and feed it into a simple Feed Forward Neural Network (FFNN) layer.

3.3 Model Performance Evaluation

We applied hyperparameter tuning via a stratified 5-fold cross-validation process on the training set to arrive at the most efficient hyperparameters. The

[7] https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter

[8] https://github.com/UBC-NLP/marbert

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

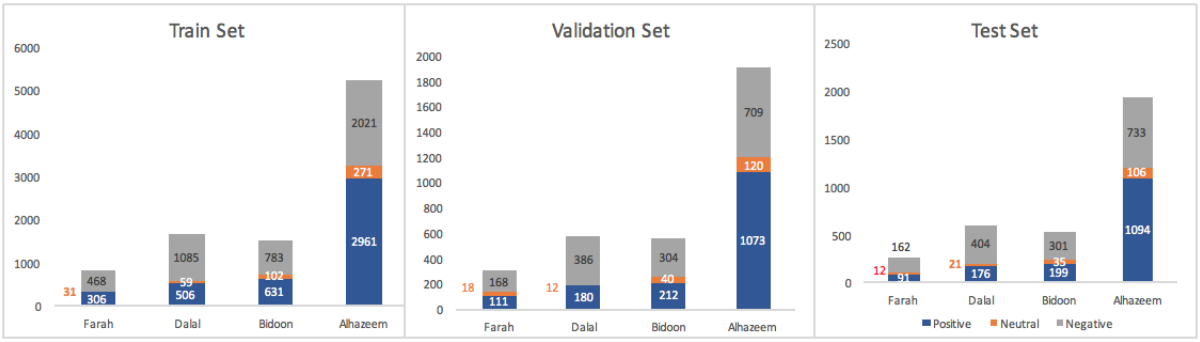

Figure 3: Classes distribution of each subset from the dataset

distribution of the sentiment classes is not equal in all sets, as can be seen from Fig.3. Thus, we depend on macro-averaged measurements to remove any bias toward a particular class. Macro F1 and accuracy were applied in most experiments. Models were evaluated using a stratified 5-fold cross-validation to remove any bias by averaging the results. The evaluation metrics were developed using the Scikit-Learn Python library. Google Colab was used to conduct the experiment. We further evaluate the results through manual inspection and error analysis.

4 Results and Discussion

Firstly, we evaluate the Snorkel annotated dataset to check for the annotation agreement between the Snorkle-labeled dataset and the gold-labeled dataset (human-labeled dataset). Thus, we consider pairwise percent agreement and Cohen's kappa coefficient metrics to evaluate annotation agreement. The result report 93% for the pairwise percent agreement and Cohen's kappa coefficient is a near-perfect agreement with a value equals to 0.87.

We also tested the performance of the Snorkel probabilistic labeling model by applying the same steps illustrated in Fig.2. Snorkel framework already provides a function to evaluate its performance in case gold labels are present in the dataset. Thus, we applied the steps to the gold-labeled testing dataset of size=1,534 tweets, and the final performance results of the labeling system were accuracy score of 93%, and F1-Macro of 84%.

Table 1 presents the results for the baseline models and Table 2 shows the results for the main classification models. As can be noticed, the SVM reports the best performance among the baseline models. However, after further training using both train and validate sets, it reports almost perfect performance with 0.99 and 1.00 for the macro-averaged F1 and accuracy scores respectively, which indicates a possibility of over-fitting. Investigating the result from the SVM model shows that only 3 positive tweets were misclassified as negative, 5 negative tweets were misclassified as positive, and for the neutral tweets, 1 tweet was misclassified as positive and 4 tweets were misclassified as negative. A similar finding is also applied to the bagging model.

The results of the BERT models highlight an important finding. Even though AraBERT includes in its pre-training dataset tweets and emoji, similar to our dataset, and MARBERT is developed using a large tweets dataset, they both were not performing as well as ARBERT. The ARBERT model reports 0.75 and 0.89 for the macro-averaged F1 and accuracy scores respectively on the test set.

	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 1: Baseline 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

Table 2: Main models results

4.1 Error Analysis

Since the dataset consists of a large number of tweets, explicit sentiment tweets and more ambiguous ones were encountered. The explicit tweets were clear, easy to classify, and convey sentiments by both Snorkel and human annotators. On the other hand, various tweets were challenging to classify. Some were not clear in terms of the focus of the topic as the reader would find the meaning complicated to understand, and others were difficult to decide their suggested sentiment. Samples from the explicit sentiment and ambiguous tweets are presented in Appendix A.

Additionally, one noted observation while going through the tweets was that they contained foreign vocabularies that were borrowed from other languages (English in most cases), modified to fit the Kuwaiti dialect, or just written in Arabic alphabets like (بريك / *break* and اوفر تايم / over time), and used regularly among Kuwaitis, showing that the Kuwaiti dialect is constantly updating with new words added to it.

5 Conclusions

In this paper, we release the first open large-scale dataset focused on sentiment analysis for the Kuwaiti dialect using a semi-supervised approach. We created a semi-supervised model based on the Snorkel framework to reduce the need for human annotators and boost the size of the labeled data rapidly and accurately. To test the applicability of the dataset, we evaluated various traditional machine learning classifier baselines, as well as advanced BERT-based language model classifiers. The results showed that our approach generates high-performance scores in both macro-average F1 and accuracy results. We believe our approach will help foster research and development of NLP systems, which were previously little studied due to the challenges faced by human annotators.

6 Future Work

To further prove the validity and significance of our proposed weak supervised labeling system, we plan to test the labeling methodology on Arabian Gulf dialects other than the Kuwaiti dialect. Furthermore, for labeling functions in Snorkel, we plan to test various versions of the prompt text used in the zero-shot pre-trained models using different Arabic and English prompts and by testing the effect of combining rule-based and heuristic labeling functions with zero-shot pre-trained models on the accuracy of weak supervised labeling system.

7 Ethics Statement

We constructed the sentiment analysis Kuwaiti dataset using the public tweets that span several time-frames and themes, Snorkel open-sourced framework for automatic labeling, and human annotators for the annotation evaluation dataset. All sensitive and personalized content was removed from the tweets for users' privacy concerns. An SME who is an expert in NLP, Kuwaiti dialect, and Snorkel framework administrated the creation of labels using Snorkel to ensure the accuracy of the automatic annotation process. We only recruited Kuwaiti annotators that are fluent Kuwaiti speakers, with a very high approved task acceptance rate to label the evaluation dataset manually.

References

Arwa A. Al Shamsi and Sherief Abdallah. 2022. Sentiment analysis of emirati dialect. *Big Data and Cognitive Computing*, 6(2).

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah 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 (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Snorkel AI. Advancing snorkel from research to production. Last accessed 02 September 2022.

Shamlan Al-Qenaie et al. 2011. *Kuwaiti Arabic: A socio-phonological perspective*. Ph.D. thesis, Durham University.

Rahma M. Alahmary, Hmood Z. Al-Dossari, and Ahmed Z. Emam. 2019. Sentiment analysis of saudi dialect using deep learning techniques. In *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–6.

Banan A. Alharbi, Mohammad A. Mezher, and Abdullah 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(6).

Meshrif Alruily and Osama R Shahin. 2020. Sentiment analysis of twitter data for saudi universities. *International Journal of Machine Learning and Computing*, 10(1).

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Joe Davison. xlm-roberta-large-xnli.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jesper E. van Engelen and Holger H. Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*.

Daniel Fu, Mayee Chen, Frederic Sala, Sarah Hooper, Kayvon Fatahalian, and Christopher Ré. 2020. Fast and three-rious: Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, pages 3280–3291. PMLR.

Nizar Y. Habash. 2010. 1 edition, volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan Claypool Publishers. [link].

Noor A Hayat and Yousuf B AlBader. 2022. The mchicken phenomenon: How has english become a prevalent language among kuwaiti youths? *World Journal of English Language*, 12(6).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv*.

Fatemah Husain. 2020. OSACT4 shared task on offensive language detection: Intensive preprocessing-based approach. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 53–60, Marseille, France. European Language Resource Association.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1).

Fatemah Husain and Ozlem Uzuner. 2022. Investigating the effect of preprocessing arabic text on offensive language and hate speech detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(4).

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Thuraya M Omran, Baraa T Sharef, Crina Grosan, and Yongmin Li. 2022. Transfer Learning and Sentiment Analysis of Bahraini Dialects Sequential Text Data using Multilingual Deep Learning Approach. *Chaos, Solitons and Fractals*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

A. Ratner, P. Varma, B. Hancock, and C. Ré. Weak supervision: A new programming paradigm for machine learning. Last accessed 02 September 2022.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29.

Sandra Rizkallah, Amir Atiya, Hossam ElDin Mahgoub, and Momen Heragy. 2018. Dialect versus msa sentiment analysis. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, pages 605–613, Cham. Springer International Publishing.

Joshua Robinson, Stefanie Jegelka, and Suvrit Sra. 2020. Strength from weakness: Fast learning using weak supervision. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8127–8136. PMLR.

Janan Ben Salamah and Aymen Elkhlifi. 2014. Microblogging opinion mining approach for kuwaiti dialect. In *The International Conference on Computing Technology and Information Management (ICCTIM)*, page 388. Citeseer.

Burr Settles. 2009. Active learning literature survey.

Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2022. Language Models in the Loop: Incorporating Prompting into Weak Supervision. *arXiv*.

W.H. Tok, A. Bahree, and S. Filipi. 2021. *Practical Weak Supervision: Doing More with Less Data*. O'Reilly Media, Incorporated.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. " O'Reilly Media, Inc.".

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

S. Yildirim and M. Asgari-Chenaghlu. 2021. *Mastering Transformers: Build state-of-the-art models from scratch with advanced natural language processing techniques*. Packt Publishing.

A Appendices

A.1 Explicit Tweets

Explicit tweets refer to tweets that contain some verbs or nouns expressing the feelings and opinions of the author clearly. These tweets were very easily classified based on the sentiment labels; negative, positive, or neutral, by the proposed Snorkel system and human annotators as well. Samples from these tweets are shown in Table 3.

A.2 Ambiguous Tweets

Ambiguous tweets refer to tweets that contain unclear text that is complicated in terms that it shows feelings and emotion but it is not clear whether this sentiment is negative, positive, or neutral. Thus, it is not factual or news, rather it illustrates some sentiment but the state if the sentiment is not stable. Table 4 shows some examples of ambiguous tweets from the dataset.

Sentiment	Tweet
Positive	الله موجود ،،، ما ورى هالشكول حلول ..! #كفى ـاستهتار ـبمصير ـالبدون *God exists,,,there are no solutions for those people enough negligence of the Bidoon*
Negative	دائما فجر تظهر في هذا الموقف لتعطي انطباع الامان للجهاز المركزي فجر أنتي من أبواق الجهاز ومن اللاتي تعتاش على قضية البدون انسانة في غاية الخسة *Fajer always appears in this position to make an impression of the safety for the central control. Fajer, you are one of the horns of the central control, and among those who subsist on the issue of the Bidoon, a very mean person*
Neutral	الخدمة المدنية.. غير صحيح ما يتداول بشأن رفض الديوان صرف مكافأة #الصفوف ـالأمامية لـ #البدون *The Civil Service.. incorrect rumors about the refusal of the Diwan to disburse a reward to the front rows of the Bidoon*

Table 3: Samples from the explicit tweets

Challenge	Tweet
Not direct. It could be sarcastic by referring to the amount of attention and empathy Rayan was getting, but could also be serious, free from sarcasm	لم نجد هذا الكم من التعاطف لأطفال البدون الذين يتساقطون يومياً في بئر الحرمان لم نجد أبداً هذا الكم من المشاعر لشباب #البدون *We did not find this amount of sympathy for the Bidoon children who fall daily into the well of deprivation. We have never found this amount of feelings for the Bidoon youth have never found this amount of feelings for the Bidoon youth*
The sentiment here was both positive and negative, as the idea of unity gave a positive feeling, but stating the issues they were facing gave a negative one.	وستبقى قضية الكويتيين #البدون ـاولويه عند كل شريف في هذا الوطن الجمعة القادمه ٢٢٠٢ الساعة الواحدة بعد صلاة الجمعة في تيماء العزة والكرامة والحرية وقفة الكويتيين البدون جنسية ضد الظلم والإهانة والتسويف والتجاهل ولكي *The issue of Kuwaitis Bedoons will remain a priority for every honorable person in this country, next Friday 11/2/2022 one o'clock after Friday prayers in Taima, the pride, dignity and freedom. The stateless Kuwaitis stand against injustice, humiliation, procrastination and disregard, and for our message to reach everyone, we are a people who deserve to live with dignity.*

Table 4: Samples from the ambiguous tweets