Social Network Analysis and Mining (2023) 13:11
https://doi.org/10.1007/s13278-022-01009-0

ORIGINAL ARTICLE

An ensemble transformer-based model for Arabic sentiment analysis

Omar Mohamed[1] · Aly M. Kassem[2] · Ali Ashraf[1] · Salma Jamal[3] · Ensaf Hussein Mohamed[1]

Received: 2 October 2022 / Revised: 24 November 2022 / Accepted: 26 November 2022

Abstract
Sentiment analysis is a common and challenging task in natural language processing (NLP). It is a widely studied area of research; it facilitates capturing public opinions about a topic, product, or service. There is much research that tackles English sentiment analysis. However, the research in the Arabic language is behind other high-resource languages. Recently, models such as bidirectional encoder representations from transformers (BERT) and generative pre-trained transformer (GPT) have been widely used in many NLP tasks; it significantly improved performance in NLP tasks, especially sentiment analysis. However, Arabic was not a priority in their development. Several models focusing on Arabic have recently begun to pave the way for the latest technologies, such as ARBERT, MARBERT, and others. We used multiple datasets for training and testing-ASAD-A Twitter-based Benchmark Arabic Sentiment Analysis Dataset, ArSarcasm-v2, and SemEval-2017. We propose an ensemble learning approach that combines the multilingual model(XLM-T) and the monolingual model(MARBERT) to overcome the intricacies of the Arabic language that are difficult to address with a single model. It also addresses the problem of imbalanced data using a combination of focal loss and label smoothing. The experiments showed that our ensemble learning approach outperforms the state-of-the-art models on all the used datasets.

Keywords NLP · Arabic text · Sentiment analysis · Ensemble learning · Transformers · BERT

1 Introduction

Over the past two decades, research on sentiment analysis (SA) and subjective linguistic analysis has emerged in natural language processing (Darwish et al. 2021; Abo et al. 2019; Oueslati et al. 2020). Sentiment analysis has become a vital tool in text classification and interpretation with the spread of social media. Sentiment analysis studies people's opinions, sentiments, emotions, and attitudes. It is usually scaled to three negative, neutral, or positive values. Most of the research on SA has focused on English, while other languages, including Arabic, have been delayed due to their complexity. Due to diacritics and complex word combinations, Arabic is a complex language that can change the meaning or sentiment (The Editors of Encyclopaedia 2021). The Arabic language contains various dialects used in daily life, which may differ from one country to another and even differ in the same country itself. Colloquial Arabic differs from Modern Standard Arabic (MSA) in many aspects because it sometimes does not follow specific grammatical rules and has different pronunciations of words. It also contains many words borrowed from other languages or specific to that dialect. The basis of the many sentiment analysis approaches is the sentiment lexicon,

Omar Mohamed, Aly M. Kassem, Ali Ashraf and Salma Jamal have contributed equally to this work.

✉ Aly M. Kassem
kassem6@uwindsor.ca

Omar Mohamed
omar_20170353@fci.helwan.edu.eg

Ali Ashraf
aliashraf@fci.helwan.edu.eg

Salma Jamal
sagamal@nu.edu.eg

Ensaf Hussein Mohamed
ensaf_hussein@fci.helwan.edu.eg

[1] Department of Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University, Helwan, Egypt

[2] School of Computer Science, University of Windsor, Windsor, Canada

[3] School of Information Technology and Computer Science, Nile University, Giza, Egypt

Published online: 24 December 2022

with words and phrases classified as conveying positive or negative sentiments (Jurek et al. 2015; Kaushik and Mishra 2014; Taboada et al. 2011). Emojis are often overlooked as a reliable source of information, yet Emojis proved to be an integral feature within the sentiment analysis process (Shiha and Ayvaz 2017; Liu et al. 2021; Khan and Peacock 2019; Rabbimov et al. 2020; Mubarak et al. 2022). However, most sentiment analysis research has focused on emoji in languages apart from Arabic and, consequently, most of the resources developed(Antoun et al. 2020; Farha and Magdy 2019; Abdelali et al. 2021). Arabic foists many provocations for NLP tasks in general; these include dialectal variation, morphological complexity, ambiguity, and lack of resources. For the past decade, advancements in contextualized language representations revolutionized NLP, beginning with pre-trained word representations such as Word2Vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014), and FastText (Bojanowski et al. 2017). However, these word representations were static and did not rely on the context in which they appeared. The tokenization process relies on naive methods in contrast with state-of-the-art model tokenizers, such as subword tokenization methods (Wu et al. 2016; Kudo 2018; Sennrich et al. 2015) that are based on the premise that frequently used words should not be broken down into smaller subwords. In contrast, unusual words should be broken down into meaningful subwords. Recently, important milestones were the introduction of Attention mechanisms (Bahdanau et al. 2015), namely, Self-Attention, Transformers (Vaswani et al. 2017), and BERT (Devlin et al. 2019).

In this research, we proposed an ensemble Arabic Sentiment Analysis (ASA) model that is built using the multilingual model (XLM-T) (Barbieri et al. 2021) and monolingual model (MARBERT) (Abdul-Mageed et al. 2021) transformer models, showing the importance and effectiveness of using Emojis in training the ASA model, also reducing the negative impact of an unbalanced dataset problem through using Focal Loss (Lin et al. 2017) combined with label smoothing (Müller et al. 2019; Szegedy et al. 2016). The significant contributions of this research can be summarized as follows:

1. Overcome the dataset imbalance problem by applying a combination of focal loss and label smoothing rather than the traditional weighted cross-entropy loss.
2. Declare the effectiveness of utilizing multilingual models with monolingual models on the Arabic Sentiment Analysis task (ArSA) performance.
3. Show the importance and effectiveness of using Emojis in training the (ArSA).

The rest of this paper is organized as follows: section 2 discusses related work, section 3 presents the Methods and Materials, section 4 presents the Results and Discussion, and section 5 concludes this work while highlighting its limitations with some recommendations for future work.

2 Related work

This section discusses previous research and applications addressing SA challenges in the Arabic language. In addition, the related work methodologies, datasets, strengths employed, and drawbacks.

2.1 Arabic sentiment analysis

Al-Ayyoub et al. (2015) collected over 120,000 Arabic phrases divided into three categories: positive, negative, and neutral, and utilized a lexicon-based method to assess true sentiment. The accuracy was 87%. In social media analysis, Soliman et al. (2014) introduced a dataset of 1846 Facebook comments from various news pages, which are unorganized, unstructured, and do not follow grammatical conventions. They proposed a lexicon-based method with words and idioms and SVM with radial basis function kernel. They got 86.86% accuracy. Alayba et al. (2017) presented a new Arabic dataset for healthcare-related opinions collected between 01/02/2016 and 31/07/2016. They used various machine learning and deep learning algorithms. The best accuracy was 91% with SVM. Heikal et al. (2018) implemented a Deep Learning-based ensemble model that combined Convolutional Neural Network (CNN) (LeCun et al. 1989) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) models. Their method did not include the feature engineering phase; instead, they used the Deep Learning model to extract features automatically; the model depends on Word2vec pre-trained word embeddings (Mikolov et al. 2013) for Arabic tweets. The model achieved an F1 score of 53.6%, outperforming the state-of-the-art models on the Arabic Sentiment Tweets Dataset (ASTD)(Nabil et al. 2015). Alayba et al. (2018) focused on the impact of integrating CNNs and LSTMs networks with word-level and five-gram character levels. Their model outperformed the prior results on three separate datasets. Al-Smadi et al. (2019) proposed two approaches for two particular problems in Arabic hotel reviews. They implemented character-level bidirectional LSTM with a conditional random field layer (Zheng et al. 2015) for the first task: extracting aspect-oriented target expressions. They applied LSTM for the second challenge, aspect sentiment polarity categorization. Their baseline outperformed the previous models by 39% and 6% in the first and second tasks, respectively. Al-Twairesh and Al-Negheimish (2019) presented a model that outperformed previous results on three separate sentiment analysis datasets and established a new state-of-the-art model. They used the ensemble technique on two types of features: surface and

deep features. The first was constructed by feature engineering, whereas the second came through generic/sentiment-specific embeddings. Oussous et al. (2020) proved that the Deep Learning models were efficient on a new Arabic Moroccan Sentiment Analysis corpus. They used intensive cleaning techniques such as stemming and normalization. A comparison of Deep and Machine Learning demonstrates the significant potential of Deep Learning in Arabic Sentiment Analysis.

2.2 Arabic sentiment analysis On ASAD

This section will present all the applied models on the ASAD dataset. The main note was that all the top-ranked participants used MARBERT to represent tweets. Most participants used similar text preprocessing techniques, such as removing unrecognizable symbols or characters that do not reflect the meaning of the text and replacing extra content in tweets with unique tokens. However, each team has its strategy for using MARBERT and other deep learning approaches. We will present a quick summary of their models below.

Wissam et al. proposed (Alamro et al. 2021), an ensemble learning model composed of five separate models; after applying score averaging, they used a label-weighted average to overcome the problem of skewed prediction distribution. Some models improved the emoji representation by replacing out-of-vocabulary(OOV) emojis with the most similar emojis. In contrast, others used various preprocessing methods, such as removing double quote marks and extending the dataset by using the ArSarcasm dataset (Farha and Magdy 2020) or employing a CNN-based classifier instead of the dense layer. They achieved a Macro F-1 score of 79.620% on Test Set-1. While Alamro et al. (2021) proposed a hybrid of CNN-LSTM architecture trained on static character embeddings, words embeddings, and the BERT model (contextualized embeddings). They used MARBERT to get the contextualized embeddings, word2vec to get the word embeddings from a large corpus of tweets, and a pre-trained character representation model (CE) (Alharbi and Lee 2020) to get the character embeddings. On Test Set-1, their approach got a Macro F-1 score of 79.014%. Alamro et al. (2021) applied to stem to the tweets first, then extended the original text by adding stems. To increase the sample size of the training set, they predicted the tweet labels from the test set and then added them to the training set. This was accomplished by running several trials and including tweets with consistent labels across trials. Their proposed model got a Macro F-1 score of 79.349% on Test Set-1. CS-UM6P team (Alamro et al. 2021) proposed a deep multi-task model with four task attention layers on top of a MARBERT encoder. They used a one-versus-all (binary classification) for each sentiment and a multi-class classification. On Test Set-1, their approach got a Macro F-1 score of 79.461%.

2.3 Arabic sentiment analysis on ArSarcasm-v2

This section will present all the applied models on the ArSarcasm-v2 dataset. El Mahdaouy et al. (2021) demonstrated the effect of joint learning between sentiment analysis and sarcasm detection through an end-to-end deep multi-task learning model(MTL). MARBERT model word embedding (Abdul-Mageed et al. 2021) is used as an encoder and attention interaction module to assign several tasks. The model exhibits promising performance with both tasks by allowing knowledge sharing between them. Song et al. (2021) fine-tuned the XLM-R (Barbieri et al. 2021) and AraBERT (Antoun et al. 2020) on the ArSarcasm-v2 (Farha et al. 2021) dataset using task-adaptive pretraining (TAP) (Gururangan et al. 2020) and knowledge distillation (Hinton et al. 2015). The stacking technique is used to acquire the final true sentiment. Abdel-Salam (Abdel-Salam 2021) used the MARBERT model to solve the ASA problem. Various preprocessing techniques were used, such as removing URLs, mentions, emails, dates, numbers, punctuation, English letters, stop words, and emojis before feeding the training tweets to the MARBERT model. The proposed method got an F1 score of 65.87%. Wadhawan (2021) aimed to remove the noise in the proposed dataset using extensive preprocessing, such as deleting HTML line breaks, unnecessary characters, and non-Arabic terms and performing Farasa segmentation (Darwish and Mubarak 2016) before feeding the tweets to AraBERTv0.2-large. Their technique achieved an F1 score of 65.31%. Gaanoun and Benelallam (2021) proposed an ensemble learning model that combines three distinct models. The first model is Gaussian Naive Bayes, which has two features: surface features like punctuation marks, quote marks, and hashtags. Sentiment features such as overall sentiment score, average sentiment score, and positive/negative sentiment word count are examples of sentiment features. MARBERT is the second model. The third model is Bi-LSTM with Mazajak embeddings (Farha and Magdy 2019). A weighted ensemble between the indicated models is used to get the predictions. Their proposed model achieved an F1 score of 64.39%.

2.4 Arabic sentiment analysis on SemEval

This section will present all the applied models on the SemEval dataset. El-Beltagy et al. (2017) used a Complement Naïve Bayes as a classifier based on the findings described in (Khalil et al. 2015). The feature vector representation comprised uni-grams and bi-gram terms with their IDF weights augmented with handcrafted lexical features, yielding an F^{PN} of 61.10%. Jabreel and Moreno (2017) introduced a rich set of features that are mainly based on the bag of words (BoW) model to get better representation for Arabic tweets, such as a bag of negated

words, syntactic features, lexical features, clustering features, and embedding features. The obtained features for a collection of Arabic tweets were then passed into a Support Vector Machine classifier. Their approach achieved an F^{PN} of 57.1%. González et al. (2017) presented a model to acquire high-level abstractions from noisy representations. They combined three Convolution Recurrent Neural Networks (CRNNs) (Zhou et al. 2002), each with its features. The first network employs out-domain embeddings and a 400-dimensional word2vec model trained on the Arabic Wikipedia corpus. The second network operates in-domain models; a word2vec model was prepared using the training corpus. The polarity of the word sequences forms the final network. The three networks' output is concatenated and used as an input feature vector for MLP, yielding an F^{PN} of 46.7%. Htait et al. (2017) presented a similarity model based on a cosine similarity measure between training set tweets and a list of the most prevalent words in positive and negative from two large corpora of Arabic tweets. Feature extraction was performed on 42 million Arabic tweets using word2vec model training. Their model obtained an F^{PN} of 46.9%. Farha and Magdy (2019) proposed a model based on CNN-LSTM and word2vec as text representation. Also, they present a word2vec model constructed using a dataset of 250 million tweets, yielding an embedding vector of 300. Finally, preprocessing techniques, such as letter normalization and removing elongation, strange characters, and URLs, are applied. Their model obtained an F^{PN} of 63%.

According to the findings of this survey, we identified some limitations. Most studies did not address the issue of data imbalance, instead relying on the standard weighted cross-entropy. However, a few addressed data-level issues by extending the dataset with oversampling techniques. Furthermore, only a study has attempted combining a monolingual model with a multilingual model for Arabic Sentiment Analysis, a step critical to our model's performance. This study aims to overcome previous limitations by employing and evaluating different loss functions to handle the problem of data imbalance better. In addition, ensemble learning with two state-of-the-art models is being used to overcome several complexities in the Arabic language.

3 Methods and Materials

In this section, we will go over the employed datasets and the critical components of the proposed method, starting with the data preprocessing techniques used, then a discussion of the pre-trained models used, the ensemble learning approach, and finally, the suggested loss function presented.

Table 1 Proposed datasets

Dataset	Training	Validation	Test	Total
ASAD (Alharbi et al. 2020)	66,230	7358	19,793	93,381
ArSarcasm-v2 (Farha et al. 2021)	11,293	1,255	3,000	15,548
SemEval (Rosenthal et al. 2017)	3,019	336	6,100	9,455

Table 2 Proposed datasets statistics

Dataset	Positive	Negative	Neutral	Total
ASAD	15,215	15,267	64,518	95,000
ArSarcasm-v2	2,755	6,298	6,495	15,548
SemEval	743	1,470	1,142	9,455

3.1 Datasets

Choosing a suitable dataset is a challenging task where high-quality annotated datasets are scarce. After an extensive survey, we chose the three datasets discussed below. These datasets incorporate key features that would help make the model as general and robust as possible, such as the variations of dialects and writing styles. Table 1 illustrates the data distributions in each dataset while Table 2 illustrates the distribution of datasets classes.

3.1.1 ASAD

ASAD (Alharbi et al. 2020) is a publicly available dataset that intends to contribute to research on Arabic natural language processing in general and Arabic sentiment classification in particular. Because of the rising interest in Arabic sentiment analysis, researchers have generated a wide range of data sets to aid in future developments in this subject. The data sets for measuring Arab sentiment come from various sources, including newspapers, reviews, and social media sites. ASAD comprises 95,000 annotated tweets, 15,215 positive tweets, 15,267 negative tweets, and 64,518 neutral tweets for 95,000 annotated tweets. The dataset was collected randomly between May 2012 and April 2020 to ensure diversity. Cleaning processes such as eliminating hashtags, user mentions, URLs, images, and videos consisting of a predefined set of inappropriate Arabic keywords were utilized to ensure high-quality tweets. To obtain the dialect variation, the percentage of tweets gathered was 36%, 31%, 22%, and 10% for Modern Arabic, Khaleeji, Hijazi, and Egyptian dialects, respectively.

3.1.2 ArSarcasm-v2

ArSarcasm-v2 (Farha et al. 2021) is an extension of the original ArSarcasm (Farha and Magdy 2020) dataset. ArSarcasm-v2 combines ArSarcasm with DAICT (Abbes et al. 2020), a corpus of 5,358 tweets published in MSA, dialectal Arabic, or a combination of the two, and some personally collected tweets. Sarcasm, sentiment, and dialect were all noted in each tweet. To ensure dialect variety, the dataset includes five dialects: 10,885 tweets from MSA; 2,981 tweets from Egypt; 966 tweets from the Gulf; 671 tweets from the Levant area; and 45 tweets from the Maghreb. The total number of tweets in the final dataset is 15,548, divided into 12,548 training and 3,000 testing tweets, 6,495 neutral tweets, 6,298 negative tweets, and 2,755 positive tweets.

3.1.3 SemEval

The final dataset was the Arabic Twitter dataset provided by SemEval-2017 (Rosenthal et al. 2017). There were several dialects in the dataset. The same term appears in several languages in distinct forms, such as suffixes and prefixes, and has various definitions. As a result, the work of classification becomes more complicated. Every tweet is assigned to three categories: positive, negative, or neutral. Furthermore, with 6,100 testing tweets and 3,555 training tweets, the training data needs to be more robust and sufficiently balanced.

3.2 Methodology

Our model is divided into three significant components: data cleaning, the BERT transformer-based models (Devlin et al. 2019), and the ensemble learning process. The workflow is going as follows: Firstly, the Bert classification head was replaced with a Multi-layer Perceptron (MLP) head with 50 hidden layers. Next, MARBERT was fine-tuned using the training data, and then XLM-T-base was fine-tuned using the same training data. Finally, the outputs of both models were passed to the ensemble module. Furthermore, we increased the size of the dataset by implementing Rough-bore Pseudo-Labeling (PL) (Arazo et al. 2020), a semi-supervised technique to add confidently predicted test data to the training data. We examined several PL variants - canonical per-batch updates, altering the loss functions. The variant that performed the best from the ensemble was used to label the test samples and then add the newly labeled data to the train set, which helped the training to converge. PL improved the difference in train and test distributions, which was a leakage in the dataset.

3.2.1 Data cleaning

The cleaning process is performed to ensure the best data quality. The BERT tokenization process inspires this cleanup module. Tweets that contain at least one of the following conditions will be cleaned up relative to their respective operations. Due to some tweets deletion from the original dataset and dropping duplicated instances, the total number of records selected is 93,381. We normalized certain words into the specific token, such as URL into "" رابط"". Furthermore, we normalized hashtags and names, which may or may not add information, but normalization will add information to the model.

For instance, neutral tweets might contain frequent hashtags or mentions. In addition, elongation (Hegazi et al. 2021) was a conspicuous problem in the dataset since most Arabic speakers tend to increase the length of a word to show emphasis, such as "هههههههه" as an expression of laughter, which increased the sequence length, time, and space complexity. To remedy this problem, we have divided all elongated words into two letters. We omitted the double quotes because MARBERT's model does not include them in its tokens (OOV). The cleaning operations are shown in Table 3. After the initial cleaning, the dataset appeared to have duplicated tweets with different labels. However, seeing this anomaly prompted us to train on two datasets, one with duplicate records and another without duplicates, and to ensemble the predictions generated by training.

3.2.2 BERT - Transformer Based Model

The transformer was a breakthrough in the field of NLP and has inspired many linguistics researchers. The transformer (Vaswani et al. 2017) is a deep learning model that considers the attention mechanism and weights the significance of each component of the input separately. This encourages researchers to develop pre-trained models using

Table 3 Cleaning Operations

Condition	Action
Mentions (@user)	Replace with "" مستخدم""
URL (www.abc.com)	Replace with "" رابط""
Hashtag (#user)	Replace with "" هاشتاج""
Digits and Non Arabic letters	Remove
Emojis and emotions	Keep
Emails (user@user.com)	Replace with "" بريد""
Elongation ("" اللللله"")	Reduce elongation ("" اله"")
Redundant Punctuation ("")	Remove
Double Punctuation ("")	Remove
Parentheses (( ))	Remove

self-supervised learning (SSL), such as the masked language model (MLM): BERT (bidirectional encoder representations from transformers) and GPT (generative pre-trained transformer) (Radford et al. 2019), which are trained on large language datasets. It is composed of two modules: an encoder and a decoder. The encoder is made up of two sub-layers. The first is a multi-head self-attention mechanism that accepts encodings and assesses their relevance compared to each other to generate encodings as an output. The second type of neural network is a feed-forward neural network, which processes each input encoding independently. The generated encodings are fed into the next encoder.

BERT (Devlin et al. 2019) is a robust transformer-based architecture that produces cutting-edge outcomes in various NLP tasks. It is a bidirectional multilayered transformer encoder. BERT input is specified directly as a token sequence of one or more sentences. In this sequence, the classification token [CLS] appears first. BERT splits the phrase into two halves for a handful of phrases squeezed together as input. At the beginning of the training, a word-of-a-kind token [SEP] is used. Then, for each token, learning embeddings are added. It indicates whether the split sentence was the first or second sentence in the packed pair.

In recent years, many researchers have contributed to building large and robust pre-trained models for the Arabic language; they were trained on large datasets of Arabic tweets. In the following sections, we will give an overview of the models used in this research:

MARBERT: It is a BERT-based model (Abdul-Mageed et al. 2021) that was trained on 1 billion Arabic tweets by randomly selecting tweets from a sizeable in-house dataset of roughly 6 billion tweets made up of 15.6 billion tokens and with a sequence length of just 128. The model was trained using the same network architecture as BERT Base (MLM) but without the next sentence prediction (NSP) component. The variation in MARBERT's training data allows for better interpretability of the variations in colloquial Arabic, which proved beneficial for sentiment analysis and sarcasm detection (Farha and Magdy 2021). We experimented with many Arabic BERT-based models, but MARBERT yielded the best results. A comparative illustration of the different pre-trained models is shown in the discussion section.

XLM-T: We trained the data with XLM-T (Barbieri et al. 2021) based on the XLM-R architecture (Conneau et al. 2020). XLM-R is a multilingual MLM trained on 2.5 TB of clean Common Crawl data in 100 languages. It provides a solid improvement over previous multilingual models such as mBERT. It performs well in various tasks such as classification, sequence labeling, and question answering. XLM-T is a pre-trained model on a large corpus of 198 million tweets(1,724 million tokens) in thirty languages, making it a viable fine-tuning model for the data.

3.2.3 Ensemble Learning Model

Ensemble learning (Opitz and Maclin 1999; Ganaie and Hu 2021; Olsson 2009) is the integration of multiple models, such as classifiers or experts, to solve a particular problem. Ensemble learning aims to improve classification and prediction or reduce the likelihood of an unfortunate selection of faulty predictions. Arabic Sentiment Analysis is challenging since it has many complexities and intricacies that are practically hard to solve with one model. Our approach utilizes two fine-tuned state-of-the-art models to ensure the best possible predictions. We began by ensembling the outputs of the MARBERT and XLM-T by taking the predictions with the most significant score from each model and repeating this process for both datasets with and without duplicate records. Finally, if an ensemble receives a high score, we use it to perform a Blending Ensemble, where we ensemble successful submissions, leading to our high final score. The Ensemble Model architecture is shown in Fig. 1.

We have tried many approaches to optimize the proposed model. We will briefly introduce them in the next sections:

I - Hyperparameter Optimization: We optimized hyperparameters such as dropout, batch size, warm-up steps, learning rate, and epochs using Weights & Biases Sweeps (Biewald 2020).

Weights & Biases Sweeps is a hyperparameter search and model optimization framework. Weights & Biases Sweeps automate hyperparameter optimization and explore possible models, efficiently sample the set of hyperparameter combinations to find promising regions, and develop intuition about our model. It enables us to test alternative hyperparameter tweaking strategies, including grid search (Shekar and Dagnew 2019), random search (Bergstra and Bengio 2012), Bayesian optimization (Snoek et al. 2015), and Hyperband (Wang et al. 2018). Figure 2 shows the hyperparameter optimization process.

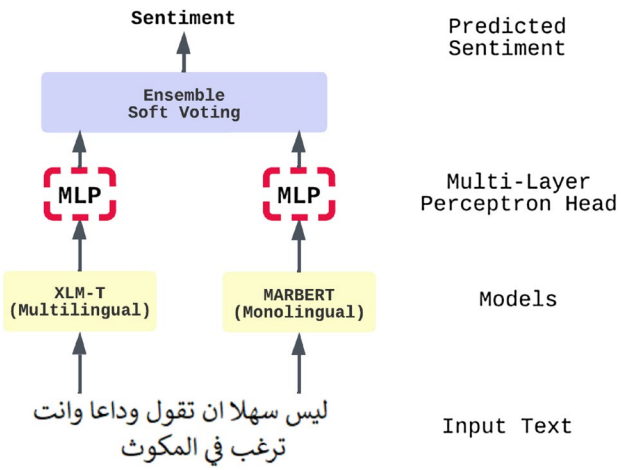


Fig. 1 Model architecture for ensemble model

II - Loss Function We used Focal Loss (Lin et al. 2017; Mukhoti et al. 2020) for the loss function. A challenging problem in the used dataset is that it is imbalanced. The focal loss function addresses class imbalance during training tasks that lack availability or diversity in their datasets, thus improving the results. The focal loss adds a modifying term to the cross-entropy loss to focus on learning about complicated negative instances. It is a dynamically scaled cross-entropy loss, with the scaling factor approaching zero as confidence in the correct class grows. This scaling factor can automatically down-weight the contribution of easy cases during training and efficiently focus the model on complex examples. The Focal Loss adds a factor $(1-p_t)^{\gamma}$ to the standard cross-entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples to ($p_t > 0.5$), emphasizing complex, misclassified cases. Here there is a tunable *focusing* parameter $\gamma \geq 0$. Also, to aid the loss function, we implemented label smoothing (Müller et al. 2019) as a regularization technique for classification problems to prevent the model from predicting the labels too confidently during training and generalizing poorly. We first tested it with weighted cross-entropy since it is one of the most common approaches for dealing with unbalanced datasets. Still, from our experiments, focal loss combined with label smoothing proved more reliable when dealing with the unbalance of our dataset. The focal loss is computed as follows:

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma} \log\left(p_t\right) \tag{1}$$

where p_t is the probability of the predicted label and γ is the focusing parameter.

III - Sampling and Augmentation There is a challenge with the used dataset; it needs to be more balanced. We tried to overcome this problem by both oversampling and downsampling. We tried augmenting the dataset by oversampling, but generating the same feature vector from the data did not increase the classification region for the same label. The new augmented data became redundant and did not benefit the model. Then, we tried augmentation by downsampling, which led to overfitting and hindering the model's performance.

4 Results and Discussion

4.1 Performance Metrics

We used many measures to assess the quality of the suggested model during the evaluation. To validate the performance of our model, we calculated Macro-F1 for both the ASAD and ArSarcasm-v2 datasets, as well as F^{PN} for the semeval dataset. F^{PN} is the traditional macro-average F-score that excludes the neutral class (Rosenthal et al. 2017). The following are the formulas for these evaluation metrics:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1_{score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

$$F1_1^{PN} = \frac{1}{2}(F_1^P + F_1^N) \tag{5}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively

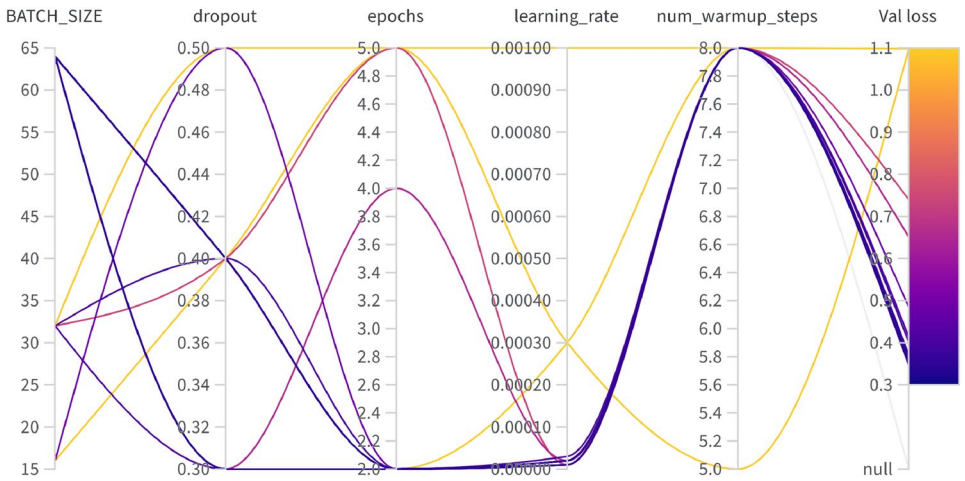


Fig. 2 Different hyperparameters value impact on validation loss

Table 4 The results of different models on ASAD (test set-1)

Model	Loss function	Classifier	Macro-F1(%)
Wissam's Approach(Alamro et al. 2021)	N/A	CNN	79.620
CS-UM6P's Approach(Alamro et al. 2021)	N/A	N/A	79.461
Ali Salhi's Approach(Alamro et al. 2021)	N/A	N/A	79.349
Abdullah I. Alharbi's Approach(Alamro et al. 2021)	N/A	Linear layer	79.014
Bi-LSMT + Mazajak	Weighted cross-entropy	Linear layer	67.180
ARBERT	Weighted cross-entropy	Linear layer	69.1
AraBERT	Weighted cross-entropy	Linear layer	70.061
MARBERT + TAP	Weighted cross-entropy	MLP (50 HS)	79.759
MARBERT	Focal Loss + Label Smoothing	Bi-LSTM	79.764
MARBERT	Focal Loss + Label Smoothing	CNN	79.776
XLM-T	Weighted cross-entropy	MLP (50 HS)	79.750
	Focal Loss + Label Smoothing		79.780
MARBERT	Weighted cross-entropy	MLP (50 HS)	79.757
	Focal Loss + Label Smoothing		79.791
Ensemble Model	Focal Loss + Label Smoothing	Linear Layer	79.909
Ensemble Model	Weighted cross-entropy	**MLP (50 HS)**	79.872
	Focal Loss + Label Smoothing		**79.986**

(N/A) Indicates that Information is Not Available

The bold usually used in papers to refer to the best performance in tables

Table 5 The Results Of Different Models On ArSarcasm-v2

Model	Loss function	Classifier	Macro-F1(%)
MTL model (El Mahdaouy et al. 2021)	Binary cross-entropy	Two linear layers	66.25
MARBERT (Abdel-Salam 2021)	Cross-entropy	N/A	65.87
Deep ensemble model (Song et al. 2021)	Cross-entropy	SVM	65.70
XLM-T	Weighted cross-entropy	MLP (50 HS)	62.25
	Focal Loss + Label Smoothing		63.10
MARBERT	Weighted cross-entropy	MLP (50 HS)	65.91
	Focal Loss + Label Smoothing		67.08
Ensemble model	Focal Loss + Label Smoothing	CNN	66.21
Ensemble model	Focal Loss + Label Smoothing	Linear Layer	66.32
Ensemble model	Focal Loss + Label Smoothing	Bi-LSTM	66.35
Ensemble Model	Weighted cross-entropy	**MLP (50 HS)**	64.17
	Focal Loss + Label Smoothing		**67.22**

(N/A) Indicates that Information is Not Available

The bold usually used in papers to refer to the best performance in tables

4.2 Experimental results

This section will discuss the results of implementing different models and architectures. To ensure that our model achieves state-of-the-art performance, we compared its effectiveness to the best-reported performance on each of the three datasets. Tables 4, 5, and 6 illustrate the results of the experiments on the three datasets. In addition, we explore all the models, classifiers, and techniques applied in experimentation, even if their results are not satisfactory.

4.2.1 Performance evaluation

The proposed model's effectiveness was evaluated by comparing it to the best models applied to the three datasets: ASAD, ArSarcasm-v2, and SemEval. The ASAD dataset, Table 4 compares the results of previous methods, different experiments,

Table 6 The Results Of Different Systems on SemEval

System	Loss function	F^{PN}
González et al. (2017)	N/A	46.7
Htait et al. (2017)	N/A	46.9
Jabreel and Moreno (2017)	N/A	57.1
El-Beltagy et al. (2017)	N/A	61.10
Farha and Magdy (2019)	N/A	63.0
Abdul-Mageed et al. (2021)	N/A	71.0
Ensemble Model	Weighted cross-entropy	69.5
	Focal Loss + Label Smoothing	**72.0**

(N/A) Indicates that Information is Not Available

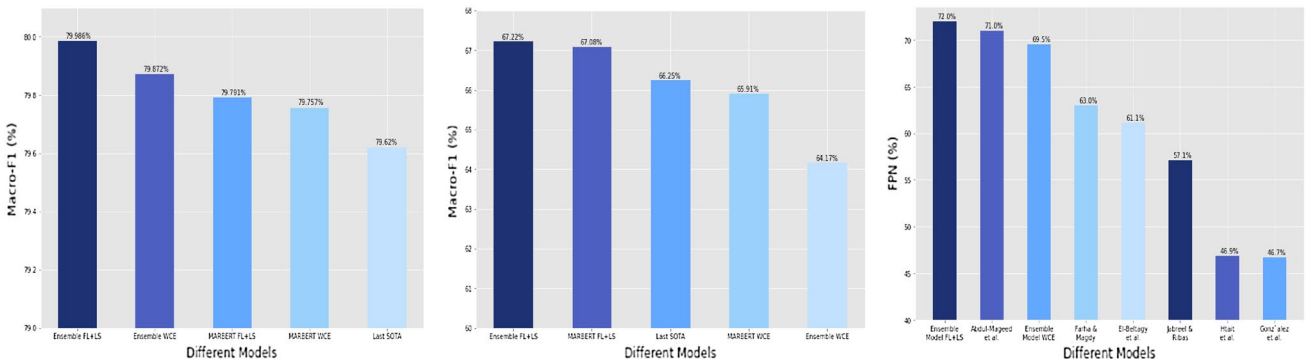
The bold usually used in papers to refer to the best performance in tables

and the proposed ensemble model. Wissam's Approach (Alamro et al. 2021) used an ensemble learning model made of five different models and achieved a macro F-1 score of 79.620%. In contrast, the proposed model outperformed it, achieving a macro F-1 score of 79.986%. The best-reported results for the ArSarcasm-v2 are by El Mahdaouy et al. (2021), who used a multi-task learning model trained jointly between sentiment analysis and sarcasm detection and achieved a macro F-1 score of 66.25%. In contrast, the proposed model outperformed by 0.97%, achieving 67.22%, as shown in Table 5. The best results for the SemEval were obtained by Abdul-Mageed et al. (2021), who applied a MARBERT model and achieved an F^{PN} of 71.0%. In contrast, the proposed model outperformed it by achieving an F^{PN} of 72.0% as shown in Table 6. The following reasons why the proposed model outperformed the prior state-of-the-art methods: the proposed model used focal loss with label smoothing to handle better the problem of data imbalance in the three corpora. Furthermore, combining multilingual models (XLM-T) with monolingual Arabic models (MARBERT) significantly improved performance for two reasons. First, because the two models generate different errors on different samples, using the ensemble learning technique was appropriate in this case because the ensemble's ability to correct the errors of some of its members is entirely dependent on the classifiers' diversity comprise the ensemble. Second, instead of using two monolingual Arabic models in the ensemble process, combine the monolingual Arabic models with a multilingual model due to the multilingual model's ability to improve performance by incorporating more training data and languages-including so-called low-resource languages, which lack extensive labeled and unlabeled datasets (Conneau et al. 2020). Figure 3 illustrates a comparison between the proposed ensemble model with different loss functions and previous state-of-the-art.

4.2.2 AraBERT and ARBERT

We tested two BERT-based models trained on Arabic datasets, AraBERT and ARBERT. AraBERT (Antoun et al. 2020) is a Bert-based model trained on 61GB of MSA text (6.5B tokens) from books, newspapers, and the dump of Arabic Wikipedia in 2019. ARBERT (Abdul-Mageed et al. 2021), a Bert-based model trained on a similar dataset of AraBERT but with a much smaller size, amounting to 24 GB of MSA text.

These two models face a related problem: Feature Learning, also known as Representation Learning (Goodfellow et al. 2016). Representation Learning essence is a combination of approaches that allows a system to automatically find the representations required for feature detection or classification from raw data. The amount of extracted information from raw data that is useful for subsequent classification or prediction is determined by data representation. The more critical the information is translated from raw data to feature representations, the better the classification or prediction performance. Despite training in cutting-edge BERT architecture, these models fail to capture the crucial properties required for sentiment analysis. Because they were trained



(a) ASAD Dataset (b) AraSarcasm Dataset (c) SemEval Dataset

Fig. 3 Comparing the proposed model with The SOTA Approaches

solely on MSA datasets, they lacked variation in dialects and could not grab the features of tweet writing style.

4.2.3 Mazajak word embedding

Mazajak (Farha and Magdy 2019) is a word2vec embedding model. It is trained on 250 million tweets gathered between 2013 and 2016 to assure subject coverage. The wide and diverse corpus ensures that many dialects are represented, lessening the impact of the dialectal variance. The model trained a Bi-LSTM model with Mazjak word embedding, but the results were less than satisfactory, yielding an F-Score of 67.018%.

4.2.4 Twitter-based pre-trained BERT models

From the experiments, we found that there is a need for a Twitter-based pre-trained model.

The first candidate was QARiB (Abdelali et al. 2021), a BERT model trained on 420 million tweets and 180 million sentences of text comprised of 14B tokens gathered using the Twitter API. Unfortunately, QARiB did not yield any impressive results since it was pre-trained without emojis included.

Experiments showed that MARBERT outperformed XLM-T individually without using the ensemble model. Nevertheless, when we ensembled the two models, the F-score was enhanced. Therefore, we used MARBERT and XLM-T as these models were trained on tweets. In addition, we trained the model on a dataset containing emojis, which dramatically improved the F-score.

We found that using MARBERT and XLM-T, pre-trained on an emoji dataset, enhanced sentence classification. For example, if a sentence includes a heart or rose emoji, the sentence most probably has a positive sentiment. It is worth mentioning that we experimented with MARBERT with TAP but have yet to yield any significant improvements over the proposed model.

4.2.5 Different Classifiers:

We tried different classifiers to ensure the best results. Namely, CNN, Bi-LSTM, and MLP. We will discuss them below:

CNN: The BERT + CNN model is based on the strategy described by (Safaya et al. 2020), which uses a CNN-based classifier instead. The following are the model steps: The text is passed through 12 layers of self-attention to create contextualized vector representations (BERT model). The contextualized embeddings of its last four hidden layers are fed into 32 filters, and convolution layers are applied. The output will then be passed to Global Max-Pooling. Finally, the output was processed through a dense layer to obtain the predictions.

Bi-LSTM: While CNN can be used to capture local text features, LSTM, which has demonstrated superior performance in text classification tasks, captures temporal information. Our experiments used two layers of Bidirectional LSTM (Bi-LSTM) with a hidden size of 128.

MLP: was used in our proposed model, which considerably enhanced the model's performance. It was easily outperformed by the MLP, outperforming the previous classifiers. MLP maximizes the use of knowledge embedded in pre-trained BERT language models by feeding the contextualized embeddings to an MLP with a hidden size of 50 and then processing the output through a dense layer to produce predictions. We began with a liner layer in our experiments instead of the MLP, but it proved to be very lacking.

The results of the different classifiers are shown in Tables 4 and 5.

Table 7 Captum Attributions (Negative -1, Neutral 0) Positive +1,

True label	Predicted label	Attribution score	Word importance
Negative	− 1 (0.77)	0.64	[CLS] الوكيل ونعم الله حسبي مستخدم مستخدم مستخدم مستخدم [SEP]
Positive	+1 (0.84)	2.76	[CLS] في الجامعه ##بات من شعور لست الصح جدا اجمل جميل يوم يناير (1).png):[SEP]
Neutral	0 (0.87)	− 1.03	[CLS] ؟ كيف الطريقة سفر جواز اصدر واي متوفي ابوي انا ، مرحبا مستخدم [SEP]

4.2.6 Inspecting The Proposed Ensemble Model

Model interpretability techniques have become increasingly important as models' complexity increases and the resulting lack of transparency. Model comprehension is a highly contentious field of study (Jacovi and Goldberg 2020; DeYoung et al. 2020; Şenel et al. 2018; Ribeiro et al. 2016; Tenney et al. 2020). We used a straightforward method for knowing which features contribute to the output of a model. We tried to open up the *Black Box* Models and provide a human-level insight into the model's inner processes. We used **Captum** (Kokhlikyan et al. 2020) to compute Integrated-Gradients for determining each word's attributions or contributions to the model decision. Also, we put our hypothesis about using emojis to see whether it helps the model to understand and predict the sentiment correctly or whether it adds some shortcut learning (Geirhos et al. 2020; Rahaman et al. 2019). The emoji *Yellow-Heart* affects the model in determining its predictions, as demonstrated in Table 7. Green indicates that the tokens are pulling toward Positive, red indicates that they are pulling toward Negative, and white indicates that they are pulling toward Neutral. The intensity of the color represents the magnitude of the signal.

5 Conclusion

In this paper, we present our novel approach to Arabic sentiment analysis. We have proposed an ensemble model that combines two models of the state-of-the-art transformer; MARBERT and XLM-T. Intensive experiments were performed on the data set. The performance of the proposed model was tested and compared with the related models. The proposed model proved its superiority with a Macro F-1 of 79.986% on the ASAD dataset, a Macro F-1 of 67.22% on the ArSarcasm-v2, and F^{PN} of 72.0% on the SemEval dataset. The inclusion of emojis in training showed their importance and effectiveness in performance. Also, adjusting the loss function has a significant effect on the model. Training the suggested model on a more extensive dataset would improve outcomes and ensure that the model is robust to noisy data and adversarial attacks (Morris et al. 2020; Xue et al. 2021). Also, applying larger models, such as the XLM-R, XL, and XXL (Goyal et al. 2021), since we have shown that combining multilingual models with monolingual models has a significant impact on model performance.

Acknowledgements This research is supported by the Vector Scholarship in Artificial Intelligence, provided through the Vector Institute.

References

Abbes I, Zaghouani W, El-Hardlo O, Ashour F (2020) DAICT: a dialectal arabic irony corpus extracted from twitter. In Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, 6265–6271. https://aclanthology.org/2020.lrec-1.768

Abdelali A, Hassan S, Mubarak H, Darwish K, Samih Y (2021) Pre-Training BERT on Arabic Tweets: Practical Considerations. arXiv preprint arXiv:2102.10684

Abdel-Salam Reem (2021) WANLP 2021 Shared-Task: Towards Irony and Sentiment Detection in Arabic Tweets using Multi-headed-LSTM-CNN-GRU and MaRBERT. In Proceedings of the Sixth Arabic Natural Language Processing Workshop. In: Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 306–311. https://aclanthology.org/2021.wanlp-1.37

Abdul-Mageed M, Elmadany A, Nagoudi E, Moatez B (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). Association for Computational Linguistics, Online, 7088–7105. https://doi.org/10.18653/v1/2021.acl-long.551

Abo MEM, Raj RG, Qazi A (2019) A review on Arabic sentiment analysis: state-of-the-art, taxonomy and open research challenges. IEEE Access 7(2019):162008–162024

Alamro H, Alshehri M, Alharbi B, Khayyat Z, Kalkatawi M, Jaber I I, Zhang X (2021) Overview of the Arabic Sentiment Analysis 2021 Competition at KAUST

Alayba AM, Palade V, England M, Iqbal R (2017) Arabic language sentiment analysis on health services, In 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR) 1, 1, 114–118. https://doi.org/10.1109/ASAR.2017.8067771

Alayba AM, Palade V, England M, Iqbal R (2018) A combined CNN and LSTM model for Arabic sentiment analysis. In: International Andreas H, Peter K, Min Tjoa A, Edgar W (eds) Machine Learning and Knowledge Extraction. Springer Publishing, Cham, pp 179–191

Alharbi AI, Lee M (2020) Combining character and word embeddings for affect in Arabic Informal social media microblogs. In: International Elisabeth M, Farid M, Helmut H, Philipp C (eds) Natural language processing and information systems. Springer Publishing, Cham, pp 213–224

Alharbi B, Alamro H, Alshehri M, Khayyat Z, Kalkatawi M, Jaber I I, Zhang X (2020) ASAD: A Twitter-based Benchmark Arabic Sentiment Analysis Dataset

Al-Twairesh N, Al-Negheimish H (2019) Surface and deep features ensemble for sentiment analysis of arabic tweets. IEEE Access 7(2019):84122–84131

Antoun Wissam, Baly Fady, Hajj Hazem (2020) AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. European Language Resource Association, Marseille, France, 9–15. https://aclanthology.org/2020.osact-1.2

Arazo E, Ortego D, Albert P, O'Connor N E, McGuinness K (2020) Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, online, 1–8

Bahdanau Dzmitry, Cho Kyunghyun, Bengio Yoshua (2015) Neural Machine Translation by Jointly Learning to Align and Translate

Barbieri F, Anke LE, Camacho-Collados J (2021) Xlm-t: a multilingual language model toolkit for twitter

Biewald L (2020) Experiment tracking with weights and biases. https://www.wandb.com/ Software available from wandb.com

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Associat Computat Linguist 5(7):135–146

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Darwish K, Habash N, Abbas M, Al-Khalifa H, Al-Natsheh HT, Bouamor H, Bouzoubaa K, Cavalli-Sforza V, El-Beltagy SR, El-Hajj W et al (2021) A panoramic survey of natural language processing in the Arab world. Commun ACM 64(4):72–81

Darwish K, Mubarak H (2016) Farasa: a new fast and accurate Arabic Word Segmenter. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, 1070–1074. https://aclanthology.org/L16-1170

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

DeYoung J, Jain S, Rajani N F, Lehman E, Xiong C, Socher R, Wallace B C (2020) ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408

El Mahdaouy A, El Mekki A, Essefar K, El Mamoun N, Berrada I, Khoumsi A (2021) Deep multi-task model for sarcasm detection and sentiment analysis in Arabic Language. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 334–339. https://aclanthology.org/2021.wanlp-1.42

El-Beltagy S R, El Kalamawy M, Soliman A B (2017) NileTMRG at SemEval-2017 Task 4: Arabic sentiment analysis. in proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, 790–795. https://doi.org/10.18653/v1/S17-2133

Farha Ibrahim Abu, Magdy Walid (2019) Mazajak: An Online Arabic Sentiment Analyser. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Florence, Italy, 192–198. https://doi.org/10.18653/v1/W19-4621

Farha Ibrahim Abu, Magdy Walid (2020) From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. European Language Resource Association, Marseille, France, 32–39. https://aclanthology.org/2020.osact-1.5

Farha Ibrahim Abu, Magdy Walid (2021) Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 21–31. https://aclanthology.org/2021.wanlp-1.3

Farha Ibrahim Abu, Zaghouani Wajdi, Magdy Walid (2021) Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 296–305. https://aclanthology.org/2021.wanlp-1.36

Gaanoun K, Benelallam I (2021) Sarcasm and sentiment detection in Arabic language a hybrid approach combining embeddings and rule-based features. In Proceedings of the Sixth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 351–356. https://aclanthology.org/2021.wanlp-1.45

Ganaie MA, Hu M et al. (2021) Ensemble deep learning: A review

González José-Ángel, Pla F, Hurtado L-F (2017) ELiRF-UPV at SemEval-2017 Task 4: Sentiment Analysis using Deep Learning. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, 723–727. https://doi.org/10.18653/v1/S17-2121

Goodfellow I, Bengio Y, Courville A(2016) Deep learning. MIT Press, online. http://www.deeplearningbook.org

Goyal N, Du J, Ott M, Anantharaman G, Conneau A (2021) Larger-Scale transformers for multilingual masked language modeling. In: Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). Association for Computational Linguistics, Online, 29–33. https://doi.org/10.18653/v1/2021.repl4nlp-1.4

Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith N A (2020) Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

Hegazi MO, Al-Dossari Y, Al-Yahy A, Al-Sumari A, Hilal A (2021) Preprocessing Arabic text on social media. Heliyon 7(2):e06191

Heikal M, Torki M, El-Makky N (2018) Sentiment analysis of Arabic tweets using deep learning. Proced Comput Sci 142(2018):114–122

Hinton G, Vinyals O, Dean J et al (2015) Distilling the knowledge in a neural network

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computat 9(8):1735–1780

Htait A, Fournier S, Bellot P (2017) LSIS at SemEval-2017 Task 4: using adapted sentiment similarity seed words for english and arabic tweet polarity classification. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, 718–722. https://doi.org/10.18653/v1/S17-2120

Jabreel M, Moreno A (2017) SiTAKA at SemEval-2017 Task 4: sentiment analysis in twitter based on a rich set of features. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, 694–699. https://doi.org/10.18653/v1/S17-2115

Jacovi A, Goldberg Y (2020) Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness?. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

James B, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(2):281–305

Jurek A, Mulvenna MD, Bi Y (2015) Improved lexicon-based sentiment analysis for social media analytics. Sec Informat 4(1):1–13

Kaushik C, Mishra A (2014) A scalable, lexicon based technique for sentiment analysis

Khalil T, Halaby A, Hammad M, El-Beltagy S R (2015) Which configuration works best? an experimental study on supervised Arabic twitter sentiment analysis. In: 2015 First International Conference on Arabic Computational Linguistics (ACLing). IEEE, online, 86–93

Khan HU, Peacock D (2019) Possible effects of emoticon and emoji on sentiment analysis web services of work organisations. Int J Work Organisat Emot 10(2):130–161

Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S et al (2020) A unified and generic model interpretability library for pytorch, Captum

Kudo Taku (2018) Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, 66–75. https://doi.org/10.18653/v1/P18-1007

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Computat 1(4):541–551

Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision. IEEE, online, 2980–2988

Liu C, Fang F, Lin X, Cai T, Tan X, Liu J, Lu X (2021) Improving sentiment analysis accuracy with emoji embedding. J Safety Sci Resil 2(4):246–252

Mahmoud A-A (2015) Essa Safa Bani, Alsmadi Izzat (2015) Lexicon-based sentiment analysis of arabic tweets. Int J Soc Network Min 2(2):101–114

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space

Mohammad A-S, Bashar T, Mahmoud A-A, Yaser J (2019) Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. Int J Mach Learn Cybernet 10(8):2163–2175

Morris J, Lifland E, Yoo J Y, Grigsby J, Jin D, Qi Y (2020) TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Online, 119–126. https://doi.org/10.18653/v1/2020.emnlp-demos.16

Mubarak H, Hassan S, Chowdhury S A (2022) Emojis as anchors to detect Arabic offensive language and hate speech

Mukhoti J, Kulharia V, Sanyal A, Golodetz S, Torr P HS, Dokania P K (2020) Calibrating deep neural networks using focal loss

Müller R, Kornblith S, Hinton G E (2019) When does label smoothing help?. In Advances in Neural Information Processing Systems, H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, and R Garnett (Eds.), Vol. 32. Curran Associates, Inc., online. https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf

Nabil M, Aly M, Atiya A (2015) ASTD: Arabic sentiment tweets dataset. In Proceedings of the 2015 conference on empirical methods in natural language processing. Association for computational linguistics, Lisbon, Portugal, 2515–2519. https://doi.org/10.18653/v1/D15-1299

Olsson F (2009) A literature survey of active machine learning in the context of natural language processing. In: SICS Technical Report. Swedish Institute of Computer Science, online, p 1–59

Opitz David, Maclin Richard (1999) Popular ensemble methods: an empirical study. J Artific Intell R 11(1999):169–198

Oueslati Oumaima, Cambria Erik, HajHmida Moez Ben, Ounelli Habib (2020) A review of sentiment analysis research in Arabic language. Future Generat Comput Syst 112(2020):408–430

Oussous A, Benjelloun F-Z, Lahcen AA, Belfkih S (2020) ASA: a framework for Arabic sentiment analysis. J Informat Sci 46(4):544–559

Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. In: proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Rabbimov I, Mporas I, Simaki V, Kobilov S (2020) Investigating the effect of emoji in opinion classification of Uzbek movie review comments. In: International Conference on Speech and Computer. Springer, online, p 435–445

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Rahaman N, Baratin A, Arpit D, Draxler F, Lin M, Hamprecht F, Bengio Y, Courville A (2019) On the spectral bias of neural networks. In: International Conference on Machine Learning. PMLR, online, p 5301–5310

Ribeiro M, Singh S, Guestrin C (2016) Why Should I Trust You?: explaining the predictions of any classifier. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics, San Diego, California, 97–101. https://doi.org/10.18653/v1/N16-3020

Robert G, Jörn-Henrik J, Claudio M, Richard Z, Wieland B, Matthias B, Wichmann Felix A (2020) Shortcut learning in deep neural networks. Nature Mach Intell 2(11):665–673

Rosenthal S, Farra N, Nakov P (2017) SemEval-2017 Task 4: sentiment analysis in twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, 502–518. https://doi.org/10.18653/v1/S17-2088

Safaya A, Abdullatif M, Yuret D (2020) KUISAIL at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media. In: Proceedings of the fourteenth workshop on semantic evaluation. International Committee for Computational Linguistics, Barcelona (online), 2054–2059. https://doi.org/10.18653/v1/2020.semeval-1.271

Sennrich R, Haddow B, Birch A (2015) Neural machine translation of rare words with subword units

Shekar BH, Dagnew G (2019) Grid search-based hyperparameter tuning and classification of microarray cancer data. In 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP). IEEE, online, 1–8

Shiha M, Ayvaz S (2017) The effects of emoji in sentiment analysis. Int J Comput Electr Eng (IJCEE) 9(1):360–369

Snoek J, Rippel O, Swersky K, Kiros R, Satish N, Sundaram N, Patwary M, Prabhat MR, Adams R (2015) Scalable bayesian optimization using deep neural networks. In International conference on machine learning. PMLR, online, 2171–2180

Soliman T-H, Elmasry MA, Hedar A, Doss MM (2014) Sentiment analysis of Arabic slang comments on facebook. Int J Comput Technol 12(5):3470–3478

Song B, Pan C, Wang S, Luo Z (2021) DeepBlueAI at WANLP-EACL2021 task 2: a deep ensemble-based method for sarcasm and sentiment detection in Arabic. In Proceedings of the Sixth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 390–394. https://aclanthology.org/2021.wanlp-1.52

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, online, p 2818–2826

Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Computat Linguist 37(2):267–307

Tenney I, Wexler J, Bastings J, Bolukbasi T, Coenen A, Gehrmann S, Jiang E, Pushkarna M, Radebaugh C, Reif E, et al (2020) The

language interpretability tool: extensible, interactive visualizations and analysis for NLP models. (2020)
The Editors of Encyclopaedia (2021) Arabic language. https://www.britannica.com/topic/Arabic-language
Utlu I, Yücesoy V, Koc A, Cukur T, Senel L-K (2018) Semantic structure and interpretability of word embeddings. IEEE/ACM Trans Audio, Speech Language Process 26(10):1769–1779
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I (2017) Attention is all you need. In advances in neural information processing systems, I Guyon, U-Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett (Eds), Vol. 30. Curran Associates, Inc., online. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
Wadhawan A(2021) Arabert and farasa segmentation based approach for sarcasm and sentiment detection in arabic tweets
Wang J, Xu J, Wang X (2018) Combination of hyperband and Bayesian optimization for hyperparameter optimization in deep learning
Wu Y, Schuster M, Chen Z, Le Q V, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K et al (2016) Google's neural machine translation system: Bridging the gap between human and machine translation
Xue L, Gao M, Chen Z, Xiong C, Xu R (2021) Robustness evaluation of transformer-based form field extractors via form attacks
Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr P HS (2015) Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. online, p 1529–1537
Zhou Z-H, Wu J, Tang W (2002) Ensembling neural networks: many could be better than all. Artific Intell 137(1–2):239–263