


Article

Fine-Grained Sentiment Analysis of Arabic COVID-19 Tweets Using BERT-Based Transformers and Dynamically Weighted Loss Function

Nora Alturayef ^{1,2,†}  and Hamzah Luqman ^{1,3,*,†} 

- ¹ Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dammam 31261, Saudi Arabia; g201902190@kfupm.edu.sa
- ² Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia
- ³ The Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
- * Correspondence: hluqman@kfupm.edu.sa; Tel.: +966-13-8601349
- † Both authors contributed equally to this work.

Abstract: The outbreak of coronavirus disease (COVID-19) has affected almost all of the countries of the world, and has had significant social and psychological effects on the population. Nowadays, social media platforms are being used for emotional self-expression towards current events, including the COVID-19 pandemic. The study of people's emotions in social media is vital to understand the effect of this pandemic on mental health, in order to protect societies. This work aims to investigate to what extent deep learning models can assist in understanding society's attitude in social media toward COVID-19 pandemic. We employ two transformer-based models for fine-grained sentiment detection of Arabic tweets, considering that more than one emotion can co-exist in the same tweet. We also show how the textual representation of emojis can boost the performance of sentiment analysis. In addition, we propose a dynamically weighted loss function (DWLF) to handle the issue of imbalanced datasets. The proposed approach has been evaluated on two datasets and the attained results demonstrate that the proposed BERT-based models with emojis replacement and DWLF technique can improve the sentiment detection of multi-dialect Arabic tweets with an F1-Micro score of 0.72.

Keywords: BERT; COVID-19 tweets; emotion detection; sentiment analysis; transformers



Citation: Alturayef, N.; Luqman, H. Fine-Grained Sentiment Analysis of Arabic COVID-19 Tweets Using BERT-Based Transformers and Dynamically Weighted Loss Function. *Appl. Sci.* **2021**, *11*, 10694. <https://doi.org/10.3390/app112210694>

Academic Editors: Alfredo Milani, Valentina Franzoni and Giulio Biondi

Received: 11 October 2021
Accepted: 10 November 2021
Published: 12 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Corona virus disease or COVID-19 was first reported by the Chinese public health authorities in the city of Wuhan in 2019 to be characterized later as a pandemic by the World Health Organization (WHO). This pandemic affected almost all world countries with more than 143 million reported cases and over 3 million deaths [1]. Many people around the world have lost their jobs during this pandemic, or have been forced to study or work remotely from home.

The study of people's feelings is vital to investigate the effect of COVID-19 pandemic on mental health. Although there are numerous studies analyzing the impacts of the pandemic on healthcare, medical treatments, and the economy, there has been relatively little emphasis on studying people's feelings during this pandemic. It is crucial to understand the personal level in order to protect societies from distress, anxiety, and mental illness.

Social media has become an inherent part of our daily life as a medium of communication. It encourages emotional self-expression toward current events including COVID-19 pandemic. Social media platforms, such as Twitter and Facebook, are considered the global center of big data, with a massive amount of generated data by people who use and spend

excessive hours on these applications [2]. This data helps in measuring people's emotions and opinions toward the COVID-19 pandemic through sentiment analysis systems.

Sentiment analysis or opinion mining is contextual mining of text which identifies and extracts subjective information from the text. It helps in analysing people's opinions and emotions toward entities such as products, individuals, and events. Sentiment analysis systems aim to identify the polarity expressed in the text [3]. These systems can be classified based on the identified text polarity into coarse-grained and fine-grained. Coarse-grained sentiment analysis techniques classify emotions broadly into three polarities: positive, negative, and neutral. Fine-grained systems deal with more sentiment classes and they can obtain more precise sentiment polarity such as sad, annoyed, official, and joking [4].

Most of the social media sentiment analysis studies follow a psychological model to annotate the training data at the finer level such as those proposed by Plutchik et al. [5], Russell et al. [6], and Ekman [7]. Plutchik's model [8] is more adopted in natural language processing (NLP) which considers sentiments as a discrete set of eight basic emotions: *joy*, *sadness*, *anger*, *fear*, *trust*, *disgust*, *anticipation*, and *surprise*. Plutchik model arranges these emotions such that opposite emotions (e.g., joy—sadness) appear opposite to each other, and emotion closer to the center have higher intensity. In addition, Plutchik hypothesized how basic emotions with varying intensities can be combined to form secondary emotions; for example, optimism as the combination of anticipation and joy [5]. However, other models may be followed to annotate the text with different sentiments [9].

Most of the current sentimental analysis studies consider the coarse-grained sentiments [8]. However, there has been an increasing interest in more informative sentiment representation by including different groups of emotional states. These fine-grained systems are more challenging, especially with the lack of labeled data at the finer level [10]. The feelings of people during the COVID-19 pandemic are more complicated, where several fine-grained sentiments can be expressed in the text. For example, people may feel *sad* and *angry* because of the rising number of deaths and losing jobs, whereas others may be *optimistic* about the updated news of the COVID-19 vaccine. Therefore, fine-grained labels are needed to better understand the people feelings during COVID-19 pandemic.

In this work, we propose a fine-grained sentiment analyzer for Arabic COVID-19 tweets by targeting 11 sentiments. We fine-tune two versions of the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT), AraBERT and MARBERT, for our task and compare them with a vanilla deep learning model. We also show the importance of emojis in reflecting the sentiment of the sentence. In addition, we propose a dynamically weighted loss function to handle the issue of imbalanced data. The proposed models have been evaluated on SenWave [9] dataset and the obtained results outperformed other techniques.

The remainder of this paper is organized as follows: Section 2 reviews most of the related works dealing with sentiment analysis of COVID-19 tweets. Section 3 describes the proposed techniques for sentiment analysis of Arabic COVID-19 tweets. The experiments and the obtained results are discussed in Section 4. Finally, Section 5 concludes the paper and highlights the contributions of the paper.

2. Literature Review

Sentiment analysis techniques can be classified into two categories: *lexicon-based* and *machine learning* techniques [11].

The lexicon-based approaches detect the emotion from the semantic polarity of words or phrases in the text [12]. These approaches depend on a predefined list of labeled instances of words or phrases to train their supervised classifiers [13]. Each instance in this lexicon will be associated with one or more emotions. For example, words such as "great" and "wonderful" are words with positive polarity, whereas "bad" and "scary" words induce negative feeling. The sentiment polarity score of the text will be computed based on this lexicon [14]. This lexicon can be generated manually or automatically using a few words as a seed to expand the lexicon lists [3]. In [15–17], the authors utilized pre-defined

lexicons, such as VADER [18], to assess the polarity score of a given text in accordance with their positive and negative values. However, creating a lexicon is a time consuming task that requires preparing large entries covering all sentiment words of a certain language.

Machine learning approaches depend on extracting features from labeled text and map them to the sentiment polarity of that text. These approaches can be classified into classical machine learning approaches and deep learning approaches. Classical machine learning classifiers, such as support vectors machine (SVM), Logistic Regression, and Naïve Bayes, predict the text's sentiment by learning from predefined features to capture different aspects of a given text. Chakraborty et al. [19] used fuzzy inference with VADER sentiment lexicon to label the data into three classes: positive, negative, and neutral. They employed hyper-parametric machine learning classifiers (Naïve Bayes, AdaBoost, and Logistic Regression) to analyze 226,668 English tweets related to COVID-19. The F1-score of their proposed model yields up to 79%. Similarly, Samuel et al. [20] and Wrycza et al. [21] used classical machine learning algorithms supported by necessary textual data visualizations to provide insights into COVID-19 sentiment progression. Samuel et al. [20] proposed a coarse-grained sentiment analysis for COVID-19 tweets using Naïve Bayes and Logistic Regression. These machine learning classifiers obtained accuracies of 91% and 75%, respectively, on a dataset consisting of 140 English labeled tweets. Naïve Bayes classifier is also used by Wrycza et al. [21] with VADER lexicon for COVID-19 sentiment analysis. Another interesting work within this category by Sattar et al. [22] investigated the COVID-19 vaccination awareness among the public via sentiment analysis and predictive modeling (i.e., Machine Learning). Unsupervised lexicon-based approaches were applied to data sets that contain 1.2 million tweets, by using the publicly available tools TextBlob and VADER, to get the sentiment of each tweet. Various classification algorithms (e.g., SVM, Random Forest, and Linear Regression) were applied to build a forecasting model classifier.

Recently, deep learning techniques have achieved a significant success in many domains, including sentiment analysis. Deep learning offers several ways of learning the text representation in supervised and unsupervised ways with the help of the hierarchy of model layers [23]. As can be observed from several review studies [24–27], the most popular deep learning models used for sentiment analysis are Convolution Neural Network (CNN) [28], Deep Belief Networks [29], Recurrent Neural Network (which includes both GRU and LSTM) [30–32], Bi-directional Recurrent Neural Network [33], and Attention-based networks [9,34].

Imran et al. [35] utilized simple deep learning model (DNN) and LSTM, with different word embeddings for sentiment analysis of COVID-19 tweets. The proposed models were trained on two English datasets (i.e., Sentiment140 [36] and Emotional-Tweet [37]) to classify the COVID-19 related tweets into negative (disgust, anger, fear, sad) or positive (joy, surprise) sentiments. Pran et al. [38] used CNN with LSTM for classifying the text into three classes: Analytical, Depressed, and Angry. The proposed technique was evaluated on a dataset consisting of 1120 Facebook comments related to COVID-19 in Bangla language and a F1-score of 0.72 was reported.

Wang et al. [39] analyzed Chinese Weibo posts by fine-tuning the BERT transformer to classify the sentiment of COVID-19 related posts. A dataset consisting of 120,000 Chinese Weibo posts was used to train and evaluate this model and a F1-score of 0.75 was reported. Another study was performed by Luo and Xu [40] to analyze restaurants' reviews posted on Yelp.com to help restaurants better understand customers' needs during the COVID-19 pandemic. The authors showed that deep learning algorithms, bidirectional LSTM, and simple Embedding with Average Pooling, outperformed classical machine learning algorithms in a sentiment prediction on a dataset consisting of 112,412 restaurant reviews.

In a recent study by Kabir and Madria [41], the authors developed two deep learning models for fine-grained sentiment analysis that were applied to a unique emotion dataset using COVID-19 tweets for categorizing 10 different emotion labels. The first model consisted of a custom Q&A RoBERTa head to extract the key phrase, which is primarily responsible for the corresponding emotion of a tweet. The second model is proposed for

emotion classification employing BiLSTM with attention layer and auxiliary features input. Their study shows how negative emotions evolved throughout the pandemic and how they grew more optimistic over time.

Few approaches have been proposed for sentiment analysis of Arabic COVID-19 tweets [9,42,43]. Two of these approaches employed classical machine learning algorithms to classify the tweets into positive, negative, and neutral sentiments [42,43]. Aljameel et al. [42] trained their model on 10,623 tweets labeled manually with positive, negative, and neutral sentiments and an F1-score of 0.84 was reported using SVM with Bigram term frequency-inverse document frequency (TF-IDF). Addawood et al. [43] built an Arabic sentiment lexicon to assign a sentiment value for each tweet in a dataset consisting of 129,391 Arabic tweets. Then, a SVM classifier is applied to report an accuracy and F1-score of 0.98 and 0.98, respectively.

Yang et al. [9] proposed a multi-label emotion classifier for Arabic COVID-19 tweets based on AraBERT model [44]. The proposed model is a BERT-based transformer model trained on the Arabic corpus. The authors reported an F1-Macro score of 0.52 on the SenWave dataset, which has been proposed in this work. The proposed dataset consists of 10,000 Arabic tweets related to COVID-19 and it is annotated with 11 emotion classes. Another multi-label emotion detection system in COVID-19 context is proposed by Mukherjee et al. [45]. The proposed system is used to study the evolution of emotions from India-specific tweets towards the COVID-19 pandemic. Two attention-based transformers, RoBERTa and BERT, have been trained on the English version of SenWave dataset and a F1-Macro score of 0.554 was reported.

A summary of the surveyed sentiment analysis systems in COVID-19 context is shown in Table 1. As shown in the table, the progress is still slow towards building sentiment analysis systems for Arabic language. This can be attributed to several challenges such as the morphological complexity of Arabic language, different dialects, and data availability. These challenges limit the application of deep learning techniques for Arabic sentiment analysis, specifically at the finer level.

Table 1. Summary of the surveyed sentiment analysis systems in COVID-19 context (Performance refers to F1-score unless indicated otherwise).

Paper	Language	Sentiments		Number of Sentiments	Methods	Data Source	Data size	Performance
		Cg *	Fg **					
Chakraborty et al. [19]	English	✓		3	Naïve Bayes, SVM, AdaBoost, LinearSVC, Logistic Regression	Twitter	226,668	0.79
Imran et al. [35]	English		✓	6	DNN, LSTM, LSTM+FastText, LSTM+Glove	Twitter	160,000	0.82
Luo and Xu [40]	English	✓		2	Gradient boosting, Random Forest, Simple embedding+ average pooling, BLSTM	Yelp	112,412	0.92
Mukherjee et al. [45]	English		✓	11	BERT, RoBERTa	Twitter	10,000	0.55
Pran et al. [38]	Bangla	✓		3	CNN, LSTM	Facebook	1120	0.72
Samuel et al. [20]	English	✓		2	Naïve Bayes, Logistic Regression	Twitter	140	Acc. 0.91
Wang et al. [39]	Chinese	✓		3	BERT	Weibo	120,000	0.75
Wrycza et al. [21]	English	✓		2	Naïve Bayes	Twitter	523,000	-

Table 1. Cont.

Paper	Language	Sentiments		Number of Sentiments	Methods	Data Source	Data size	Performance
		Cg *	Fg **					
Yang et al. [9]	English, Spanish, French, Italian, Arabic, Chinese		✓	11	XLNet, Bert, AraBERT, ERNIE	Twitter	10,000	0.52
Addawood et al. [43]	Arabic	✓		3	Naïve Bayes, SVM	Twitter	12,939	0.98
Aljameel et al. [42]	Arabic	✓		3	Naïve Bayes, SVM, KNN	Twitter	10,623	0.84
Kabir and Madria [41]	English		✓	10	RoBERTa, BiLSTM	Twitter	10,000	0.63
Sattar et al. [22]	English	✓		3	SVM, KNN, Random Forest, Linear regression, M5 Tree	Twitter	1.2 million not labeled	NA

* Cg: Coarse-grained, ** Fg: Fine-grained.

3. Methodology

This section describes the proposed models for sentiment analysis of COVID-19 tweets. The system starts by preprocessing the data to prepare it for sentiment detection. This stage is followed by representing the input text using word embedding techniques. Then, three models are used for emotion learning and classification. This stage also involves data augmentation to handle the imbalanced data problem.

3.1. Preprocessing

Data preprocessing is an essential step for several NLP applications. This step varies slightly from one NLP application to another based on the application requirements and goals. The need for data preprocessing increases in morphologically rich languages such as Arabic and Turkish languages. Arabic text has several characteristics that make it more challenging for NLP systems such as diacritics and mixture of dialect, modern, and classical texts. These challenges can be observed obviously in social media texts [46]. Preprocessing the text before feeding it to machine learning algorithms is important and can improve the accuracy of these models sharply [47].

Several preprocessing steps have been applied to the Arabic microblogs used to train and evaluate the proposed models, as shown in Figure 1. These steps involve text cleaning, normalization, removing diacritics, removing or replacing emojis, tokenization, and removing stop words. These steps are widely used in Arabic sentiment analysis and have already shown their efficiency for several NLP tasks [48,49]. To ensure a fair comparison between the proposed models, the same preprocessing steps were applied for each model with all datasets.

The first step of the data preprocessing stage is text cleaning, which includes removing the URLs, mentions (@username), HTML, line breaks, and extra white spaces. This step is followed by normalizing Arabic letters to unify the different forms of Arabic letters such as ل , ل , ل . This step also involves normalizing the repeated characters which is important to handle the non-standard way of writing some Arabic words in the social media, such as writing مستحيل "impossible" word as مستحييييييل . Then we expanded the normalization process to remove diacritics (Tashkeel) and punctuation and normalizing elongation (Tatweel).

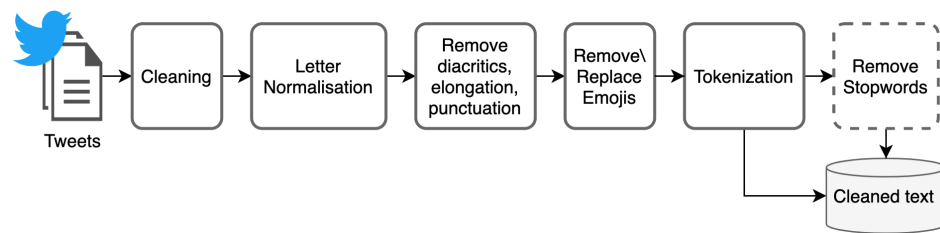


Figure 1. Preprocessing steps.

After normalizing the Arabic tweets, we proceed with text tokenization step and stop-words removal. Text tokenization is the process of segmenting the text into tokens to make it more convenient for producing words vectors. This step is followed by stop-words removal. Stop-words that do not express any emotion, such as preposition, were removed. During our experiments, we observed that removing stop-words in static-embedding based models improves the model performance. However, removing stop-words is not performed in neural models that rely on contextual embeddings, since these models consider all stop-words to provide enough context information. In addition, stop-words receive as much attention as non-stop-words in BERT-based models [50]. Lastly, to study the effect of emojis in learning process, we removed the emojis for one experiment and replaced them with their representation in another experiment, as will be discussed in Section 3.2.

3.2. Emojis Replacement

Emojis replacement is an important preprocessing step for social media text. Emojis have become widespread in social media communications, and are particularly prominent feature of expressing emotions that are difficult to express textually. For example, consider the tweets with different emojis shown in Figure 2. We can get from the first tweet that the user is joking about the situation based on the joking emoji at the end of this sentence, whereas the second tweet has anger emotion that indicates that the user is unhappy about the same situation. Thus, in this case, replacing emojis with their equivalent description could have a significant improvement in extracting the sentiment of the sentence.

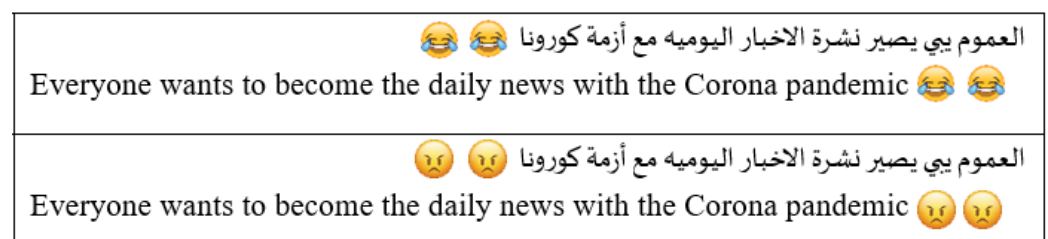


Figure 2. Two Arabic tweets with different emojis.

To incorporate the emojis into the proposed deep learning models, we represented them textually in the input tweets through emojis replacement. We started by selecting the most commonly used emojis in Arabic tweets (506 emojis) that have a meaning in emotions expression. Then, we replaced each emoji with its Arabic equivalent and represented it as a token in the input tweet.

3.3. Word Embeddings

The traditional techniques of words representation ignore the word context that degrades the accuracy of sentiment classification systems [51]. Several techniques have been proposed for word representation, such as N-gram, TF-IDF, and word embeddings [27]. Recently, pre-trained vectors (or embeddings) play a vital role in improving the accuracy of machine learning models in several problems related to the NLP [52].

Word embeddings are used extensively in NLP to capture the semantic relations between sentence's words. Embedding techniques can be classified into static word embeddings (e.g., word2vec, Glove) and contextual embeddings (e.g., BERT, ELMo). Static

word embeddings encode the context of the word into the word vector (embedding), while contextual embedding considers the sequence of all words in the documents to learn sequence-level semantics. Therefore, contextual techniques learn different representations for polysemous words [53]. In this work, we utilized both embedding representations for Arabic sentiment analysis.

Static word embedding learns a continuous representation for each word, such that each word is associated with exactly one dense vector. This word representation method was originally coined by Bengio et al. [54] to build a model that learns a distributed representation for each word along with the probability function for word sequences. The classic model of static word embedding consists of one embedding layer that feeds forward into a neural network that predicts the next word in a sequence. However, it was Mikolov et al. [55] who brought word embedding to the forefront of deep learning models for NLP, through the creation of the *Word2vec* model. In this work, we used the *AraVec* word embedding technique to represent the words in Arabic tweets [56]. This method provides powerful Arabic pre-trained word embedding models following the *Word2vec* approach. *AraVec* has several word embedding models built from three Arabic content domains (web pages, Wikipedia articles, and Twitter tweets) with two methods, Skip-Gram and Common Bag Of Words (CBOW). We used in this work the Twitter-skip-gram model with 300 dimensions (Twt-SG-300) where each word in a tweet is represented as a 300 dimensional dense vector.

We also employed contextual embedding for the sentiment analysis of COVID-19 tweets through pre-trained BERT-based transformers. The main advantage of BERT-based models is that they train word embedding based on a bidirectional transformer (or auto-encoder) rather than language model. Bidirectional transformer considers both the previous and next tokens when predicting the token, in contrast to the N-gram language model, which considers only the previous n words. Thus, the bidirectional transformers are able to incorporate contextual information from both directions at the same time. Two versions of BERT have been used in this work, AraBERT [44] and MARBERT [57]. These embeddings retrained the BERT transformer on Arabic texts. More details about these two models will be presented in the following Section.

3.4. Sentiment Analysis Models

Several deep learning models have been proposed in the literature for sentiment analysis, as discussed in Section 2. However, these models have several limitations such as ignoring the context of the text and the parallelization problem. The parallelization problem has been addressed by transformer-based models where the network depends on self-attention, allowing the model to be trained faster on more data as the implementation can be parallelized. Furthermore, transformer models (e.g., BERT) have better performance and speed in many NLP tasks [52]. In this work, we propose two BERT-based language models for emotion detection of Arabic COVID-19 tweets. These models are employed with different configurations, as will be discussed in Section 4. To evaluate the performance of these models, we proposed a baseline model consisting of a CNN model.

3.4.1. Baseline

We started by proposing a CNN model, as a baseline, to detect the sentiment of Arabic tweets and to be used as a baseline for our experiments. CNN is a special type of feed-forward neural network originally utilized in the field of computer vision [58]. Afterward, CNN showed success in NLP tasks, specifically in the classification problems [59], due to its ability in extracting important features that assist with the classification tasks.

The proposed model consists of 11 layers as shown in Figure 3. Each convolution layer applies a set of kernels to extract the features from the input word vectors. We used kernels of size 5×5 with the 'same' padding and stride of 1×1 selected empirically. The first convolution layer accepts the words' vector embeddings resulting from the text representation phase. The input to this layer is $n \times 300$, where n is the number of words

in the tweet, and 300 is the vector embedding dimension of each word. Each convolution layer in this model is followed by a nonlinear ReLU activation function. This function takes an input of a real-valued number and threshold it at zero when it is less than 0. Compared with other activation functions, such as Tanh and Sigmoid, ReLU is fast to converge and easier to compute with better performance [60]. In addition, global maximum pooling layers of size 3 are used to down-sample the obtained features of each convolution layer, as shown in Figure 3.

To overcome the overfitting problem, we regularized the network by dropping out 30% of the learned features at the dropout layer. This layer is followed by a fully-connected layer and a Sigmoid classifier. This classification layer outputs 11 output units that correspond to the number of the tweets' emotions that will be predicted by the model. We avoided using the Softmax function since it forces the total predicted probabilities to sum to one that violates our objective of predicting multiple classes or no class.

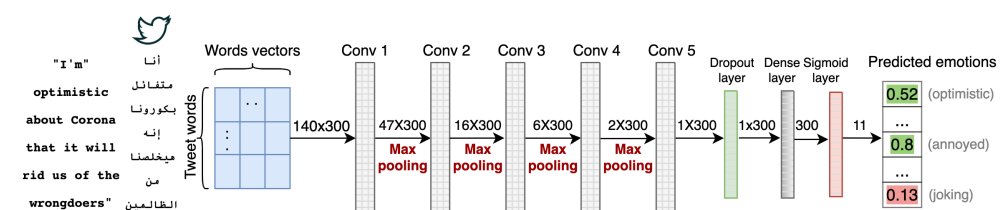


Figure 3. The architecture of the proposed CNN model for sentiment analysis.

3.4.2. BERT-Based Transformers

In this work, we employed AraBERT [44] and MARBERT [57] transformers for emotion detection of Arabic COVID-19 tweets. The two transformers are inspired by Google's BERT architecture, which is re-trained on Arabic texts. The two models use the same architecture as the BERT transformer, which has 12 attention layers, 12 attention heads, 768 hidden dimensions, and a 512 maximum sequence length. Table 2 shows a comparison between AraBERT and MARBERT in terms of data source and training parameters.

Table 2. Comparison between AraBERT and MARBERT transformers.

Model	Data Source	Vocabulary Size	#Tokens	#Parameters
AraBERT	Wikipedia, 1.5B words of Arabic corpus, OSIAN corpus, Assafir news articles	64K	2.5B	135M
MARBERT	Arabic tweets	2.5B	15.6B	163M

AraBERT and MARBERT transformers have been trained on datasets not related to the domain of COVID-19. Therefore, we fine-tuned these two transformers on datasets related to COVID-19 context. Figure 4 shows the framework of the proposed model. The system starts by tokenizing the input tweets, using WordPiece tokenizer [61], to split the word into tokens compatible with BERT-based models. Then, an input of size 128 tokens is formed and fed into the BERT-based model, AraBERT or MARBERT, to produce representations of the words in the tweets via multiple transformer layers. After representing the tweets' words using AraBERT or MARBERT models, we fed them into the classification model. As illustrated in Figure 4, only the first head of the final layer, which is corresponding to the embedding of [CLS] token, is fed into the classifier.

The classification model is a randomly initialized feed-forward layer along with a Sigmoid function to get the probability distribution over the predicted output classes. The classifier and the pre-trained model weights are trained jointly during the fine-tuning to maximize the probability of the correct sentiment. Then, we used adaptive moment estimation (AdamW) [62] for the optimization with a learning rate of $2e-5$ selected empirically.

The AdamW optimizer can be generalized better than the Adam optimizer and the models trained with AdamW optimizer have a lower training loss compared with the models trained with Adam [62]. In addition, a binary Cross-Entropy loss function was employed for the multi-label classification.

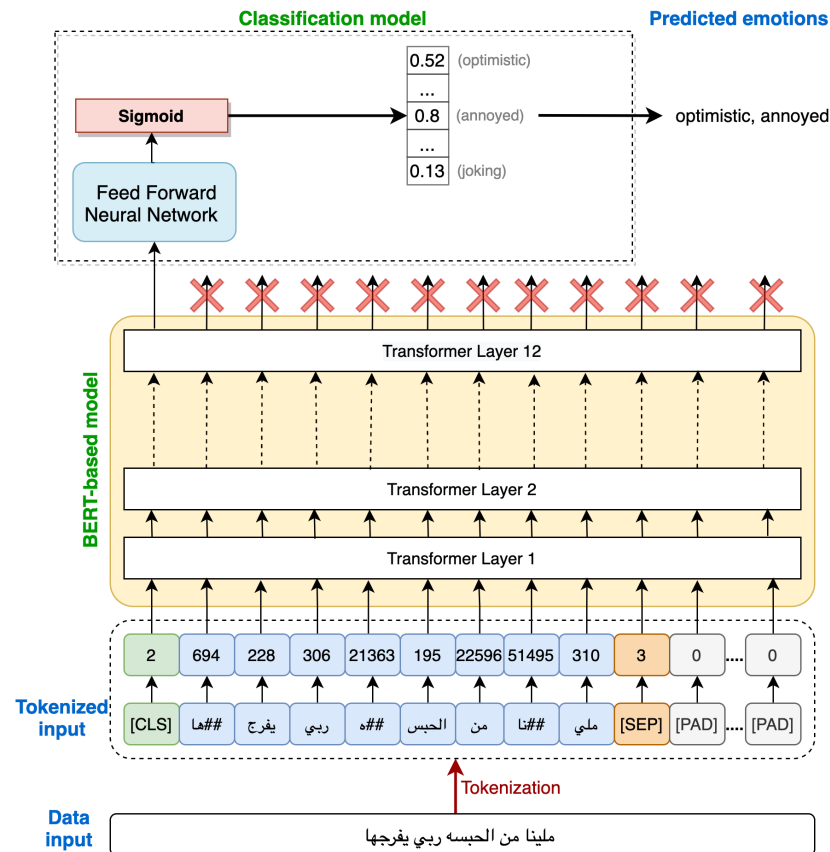


Figure 4. Framework of the proposed BERT-based model.

3.5. Data Augmentation

In multi-class and multi-label classification, a balanced dataset has target classes that are evenly distributed. If one or more classes have overwhelmingly more samples than another (i.e., there is a skewness towards these classes), this dataset can be considered as an imbalanced dataset. One of the well-known methods for handling an imbalanced dataset is to perform *oversampling* for the minority classes (classes with low samples) or *undersampling* for the majority classes (classes with a large number of samples). Although either of these two approaches balances out the dataset, they do not directly tackle this problem, rather they may introduce new issues. Oversampling entails duplication of samples associated with the minority classes. Thus, it could lead to overfitting and slow down the training process. On the other hand, undersampling the majority classes removes a certain number of samples, which could lead to the model to be disadvantaged when learning important concepts that could be learned from the removed samples.

In this work, we used the SenWave dataset [9] to train and evaluate the proposed models. To our knowledge, the SenWave dataset is the largest available fine-grained labeled dataset related to COVID-19. The dataset consists of 10,000 tweets available in 2 languages Arabic and English. In this work, we consider only the Arabic tweets. Each tweet in this dataset is labeled by 1 or more emotions, from a total of 11 emotions: *annoyed*, *anxious*, *denial*, *empathetic*, *joking*, *official*, *optimistic*, *pessimistic*, *sad*, *surprise*, and *thankful*.

The main issue of SenWave dataset is that it is an imbalanced dataset, as shown in Figure 5. Each of the *anxious*, *denial*, *empathetic*, *optimistic*, and *sad* classes (emotions) have few tweets compared with other classes. To address this issue, we proposed two

approaches. The first approach depends on creating new dataset, SenAIT, by merging the common classes of SemWave and AIT [63] datasets. The second approach is to use a dynamically weighted loss function (DWLF) that gives more weights for undersampled classes in the loss function during model training.

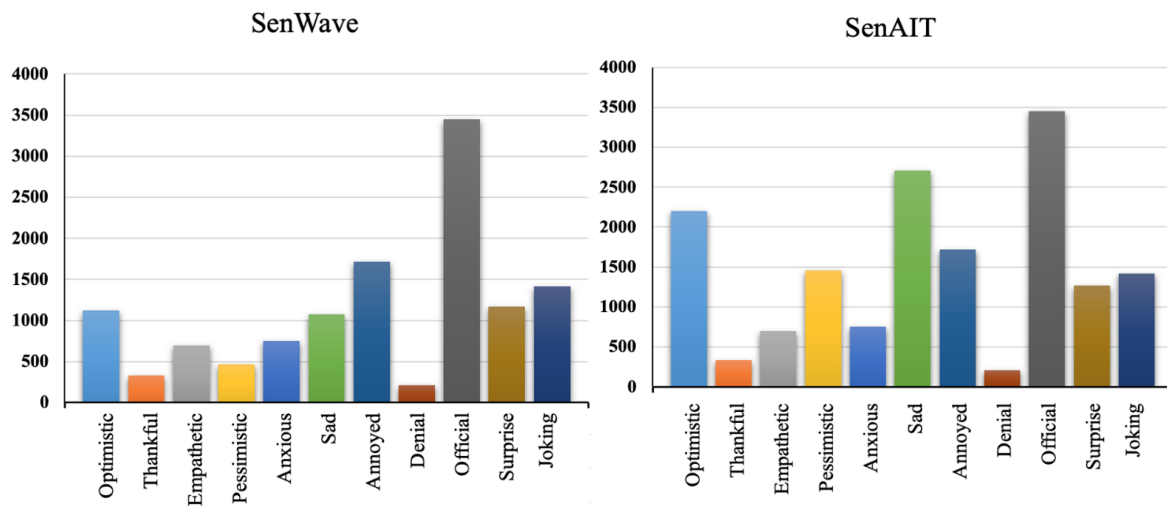


Figure 5. Emotions distribution of SenWave and SenAIT datasets.

- *SenAIT*: This dataset is created by merging the shared classes between SemWave and AIT datasets. Affect in Tweets (AIT) dataset was created as part of SemEval-2018 Task 1 [63] and it consists of 4380 Arabic tweets not related to COVID-19. Each tweet is labeled by 1 or more emotions, from a total of 11 emotions: *anger, anticipation, disgust, fear, joy, love, optimistic, pessimistic, sad, surprise, and trust*. SenWave and AIT datasets have four emotions as the common classes: *sad, surprise, optimistic, and pessimistic*. Thus, to enrich the SenWave dataset and for a comparative study, a new dataset, *SenAIT*, is created. This dataset contains all SenWave's tweets enriched by tweets from AIT dataset that are labeled by one or more of the four common classes. The resulting dataset contains 13,019 tweets with 11 classes and the distribution of these classes are shown in Figure 5.
- *Dynamically weighted loss function (DWLF)*: Another technique is proposed in this work to address the issue of imbalanced dataset through a weighted loss function. This technique involves having different weights for each class in the loss function based on the number of class's samples [64]. We assign a higher weight to the loss of the samples that belong to minor classes. For our multi-label classification task, we applied the proposed technique on the binary Cross-Entropy loss function that can be expressed mathematically as:

$$L(x, y) = \{l_1, \dots, l_N\}^T \quad (1)$$

$$l_i = -w_i [y_i \cdot \log x_i + (1 - y_i) \cdot \log(1 - x_i)] \quad (2)$$

where x_i is the input, y_i is the ground truth label, N is the batch size, and w_i is the sample weight that will optimize the contribution of the sample towards the overall loss. The sample weight is computed using the inverse of number of samples. First, the class weight w_c is computed as follows:

$$w_c = \frac{1}{\text{Number of samples in class } c} \quad (3)$$

Then, each sample weight w_i is computed as the average of weights of classes that the sample belongs to:

$$w_i = \frac{\sum_{j=1}^m w_j}{m} = 1 \quad (4)$$

where m is the number of classes that the sample belong to.

4. Experiments and Results

4.1. Datasets

Annotated data of Arabic tweets is important to build a system for Arabic sentiment analysis in the context of COVID-19. To our knowledge, SenWave is the only available dataset that can be used for emotion detection of Arabic COVID-19 tweets. This dataset consists of 10,000 tweets collected from 1 March 2020 to 15 May 2020. Each tweet is labeled by 1 or more emotions, from a total of 11 emotions. However, this dataset is highly imbalanced and we generated another dataset, SenAIT, by combining the common classes of SenWave with AIT dataset. These two datasets are used to train and evaluate the proposed models. Each dataset is divided into 80% for training and 20% for testing. More information about these datasets and the augmentation techniques used to enrich them can be found in Section 3.5.

4.2. Evaluation Metrics

Most of the state-of-the-art systems on sentiment analysis use four measuring parameters for performance evaluation: *Accuracy*, *F1 score*, *Precision*, and *Recall* [27]. However, as we have a fine-grained sentiment analysis task (multi-label classification), and to handle imbalanced data, we will consider measuring metrics that deal with multi-label problems. Therefore, we will evaluate the proposed models using six multi-label performance measures. These metrics are Multi-label accuracy, Jaccard accuracy, Macro-averaged F1 score, Micro-averaged F1 score, Label ranking average precision score, and Hamming loss. These metrics treat the data as a collection of classes and extend a binary metric (e.g., Accuracy, F1 score, and Precision) to a multi-class by averaging binary metric calculations across the set of classes [65]. These performance measures are computed as follows:

- Multi-label accuracy (*Multi-Acc.*):

$$Multi - Acc. = \frac{1}{D * m} \sum_{i=1}^D \sum_{j=1}^m \alpha(\hat{y}_{ij} == y_{ij}) \quad (5)$$

where D is the number of testing samples, m is the number of labels, y_{ij} is the ground truth label, and \hat{y}_{ij} is the predicted label.

- Jaccard accuracy (*Jac-Acc.*):

$$Jac - Acc. = \frac{1}{|D|} \sum_{i=1}^D \frac{Y_i \cap \hat{Y}_i}{Y_i \cup \hat{Y}_i} \quad (6)$$

where D is the number of testing samples, Y_i is the ground truth label, and \hat{Y}_i is the predicted label.

- Macro-averaged F1 score (*F1-Macro*):
F1-Macro = F1 score averaging on each label.
- Micro-averaged F1 score (*F1-Micro*):
F1-Micro = F1 score averaging on the prediction matrix (global calculation).
- Label ranking average precision score (*LRAP*):
LRAP = Average over each ground truth label assigned to each sample, of the ratio of true versus total labels with lower score. The goal of this metric is to assign better rank to the labels associated to each sample, and then compute whether the percentage of the higher-ranked labels were true labels.
- Hamming loss (*H-Loss*):

H -loss = The fraction of the wrong labels to the total number of labels.

4.3. Results and Discussion

Several experiments have been conducted to evaluate the proposed models on SenWave and SenAIT datasets. Table 3 shows the results of the proposed models. We compare the performance of each model (Baseline, AraBERT, and MARBERT) using SenWave dataset. Then, we evaluate these models using the newly generated dataset, SenAIT. In addition, we show how emojis replacement and DWLF approaches can improve the performance of the proposed models.

Table 3. Performance of the proposed models.

	Dataset	Method	Multi-Acc.	Jac-Acc.	F1-Micro	F1-Macro	LRAP	H-Loss	
Without DWLF	SenWave	Baseline	0.888	0.470	0.456	0.288	0.647	0.104	
		AraBERT	0.925	0.586	0.629	0.507	0.633	0.075	
		MARBERT	0.933	0.624	0.662	0.453	0.669	0.067	
	SenAIT	Baseline	0.886	0.508	0.545	0.384	0.601	0.093	
		AraBERT	0.931	0.591	0.648	0.484	0.640	0.069	
		MARBERT	0.933	0.630	0.669	0.483	0.674	0.067	
	With Emojis	SenWave	Baseline	0.892	0.476	0.502	0.289	0.655	0.095
			AraBERT	0.926	0.601	0.641	0.521	0.633	0.073
			MARBERT	0.934	0.633	0.664	0.523	0.676	0.068
	SenAIT	Baseline	0.887	0.511	0.552	0.391	0.681	0.984	
		AraBERT	0.932	0.622	0.663	0.565	0.668	0.067	
		MARBERT	0.932	0.631	0.692	0.512	0.676	0.066	
With DWLF	SenWave	Baseline	0.886	0.456	0.454	0.354	0.645	0.166	
		AraBERT	0.925	0.586	0.629	0.507	0.633	0.075	
		MARBERT	0.929	0.627	0.657	0.519	0.670	0.071	
	SenAIT	Baseline	0.888	0.460	0.465	0.396	0.657	0.171	
		AraBERT	0.931	0.591	0.648	0.484	0.640	0.069	
		MARBERT	0.930	0.632	0.671	0.568	0.673	0.070	
	With Emojis	SenWave	Baseline	0.886	0.455	0.461	0.350	0.638	0.212
			AraBERT	0.929	0.622	0.658	0.536	0.664	0.071
			MARBERT	0.931	0.638	0.665	0.530	0.680	0.069
	SenAIT	Baseline	0.887	0.477	0.480	0.400	0.678	0.163	
		AraBERT	0.936	0.651	0.694	0.632	0.691	0.064	
		MARBERT	0.932	0.636	0.725	0.704	0.678	0.068	

As shown in Table 3, the baseline model performed poorly compared with the transformer-based models, AraBERT and MARBERT, under all settings. The significant improvement of these models over the baseline model can be attributed to the contextual representation of the input tweets learned by AraBERT and MARBERT models from the large Arabic corpora used to train these models. Thus, the learned vocabulary would be more representative of Arabic morphemes and sub-word tokens, which would enable the model to learn a better contextual representation of Arabic words. In addition, we observed that the baseline model performs better for short tweets (length of 5–10 words), whereas AraBERT and MARBERT work well for longer tweets since these models are developed to capture contextual information of long texts.

Although AraBERT and MARBERT models are using the same architecture as the BERT transformer, MARBERT outperformed AraBERT on both datasets. The improvement in F1-Micro with MARBERT is between 2.0% and 3.3% compared to AraBERT. This improvement can be attributed to the fact that MARBERT utilizes massive amounts of user generated content sourced from Twitter to train the language model. These texts represent social media text characteristics, including wrong spellings, irregular syntax, abbreviations, etc., which is similar to the nature of our training datasets. In addition, these texts represent

dialectal Arabic and Modern Standard Arabic (MSA). Thus, enabling MARBERT to capture Dialectal Arabic used with the majority of tweets, whereas AraBERT was trained only on MSA.

Enriching the minority classes of the SenWave dataset with samples from AIT dataset improved the performance of the proposed models sharply as can be seen in Table 3. All models have been trained and evaluated on the SenAIT dataset resulting from this integration. As shown in the table, the improvement of F1-Micro ranges from around 1.5% to 8.9% and it can be seen with almost all experiment's arguments (emojis and DWLF).

Emojis play a significant role in expressing sentiments that cannot be expressed textually. As shown in Table 3, replacing the emojis with their textual description improved the performance across all models and datasets. For example, the F1-Micro increased by a substantial margin, 0.07–5.4%, compared with the performance of the same models without emojis. The highest improvement of F1-Micro is around 5.4% with the MARBERT model. This indicates that emojis information is complementary to the textual content, and modeling emojis by their meaning is important for emotion detection.

Although SenAIT helped in improving the performance of the proposed models, this dataset still has the issue of imbalanced classes, as discussed in Section 3.5. To address this issue, we proposed the DWLF technique, which gives more weight for the samples of the minority classes during the model's training (more information about this technique can be found in Section 3.5). As shown in Table 3, this technique boosted the performance of the AraBERT and MARBERT models sharply on both datasets. In addition, the proposed technique shows a significant improvement in detecting the minority classes (e.g., *Denial*, *Thankful*). This is supported by the confusion matrices of the minority classes before and after applying this technique. Figure 6 shows the confusion matrices of three minority classes (*Empathetic*, *Thankful*, and *Denial*) before and after using DWLF with MARBERT model. As shown in the figure, the model has improved when detecting emotions that have very limited examples in the training set.

Empathetic		Thankful		Denial	
TP	FP	TP	FP	TP	FP
61	25	12	1	0	0
FN	TN	FN	TN	FN	TN
68	2450	50	2541	41	2563
TP	FP	TP	FP	TP	FP
71	40	31	33	12	11
FN	TN	FN	TN	FN	TN
58	2435	31	2509	34	2547

Figure 6. Confusion matrices of three minority classes before (first row) and after (second row) applying DWLF technique with MARBERT model.

Furthermore, we analyzed the output of the best model, MARBERT, with emojis replacement and DWLF. The analysis aims to better understand the source of errors for tweets' emotions that were incorrectly categorized. Table 4 shows examples of the misclassified tweets from the SenWave dataset. The manual inspection of these misclassified tweets indicates the following possible source of errors:

- Unrelated tweets: It has been noticed that there are many tweets in the dataset that are not expressing any clear emotion, such as tweets #1, #2, and #3 in Table 4.
- Ambiguous emotions: There are some tweets that appear hard to have an agreement for emotion labeling. For example, tweet #4 could be labeled as "surprise" or "optimistic". Similarly, tweet #5 could be labeled as "annoyed" or both "annoyed" and "sad".

- Annotation error: Some tweets have been incorrectly annotated by the human annotators. For example, tweets #6, #7, and #8 of the SenWave dataset are annotated with unrelated emotions.

Table 4. Misclassified tweets from SenWave dataset.

ID	Tweet	Actual (Human Annotation)	Prediction (Our Method)
1	جنود الله.. إن لله جنوداً يحفظونك ويدافعون عنك، منها: عملك الصالح Soldiers of God.. God has soldiers who protect you and defend you, including: your good deeds	Empathetic	Empathetic, Thankful
2	عندي سؤال: باللغة العربية يوجد هناك ضمير غائب، فكيف للغائب ضمير؟ I have a question: In Arabic there is an absent pronoun, so how does the absent have a pronoun?	Joking	None
3	تطبيق موعد Appointment application	Official, Surprice	Official
4	بسبب كورونا اول مرة في حياتي اشترى ملابس اون لاين.يارب تطلع القياسات صح Because of Corona, for the first time in my life, I buy clothes online. I hope the size is right	Surprice	Optimistic
5	والله ما مضايقتني اكثر من غير الإشاعات الي بتطلع! بكفي!!! I swear that what annoys me more is the rumors spreading! Enough!!!	Annoyed	Annoyed, Sad
6	صباح الحرية لكل مظلوم خلف القضبان Morning freedom for every oppressed behind bars	Sad	Optimistic
7	الحشد الشعبي يطلق عملية كبح الجائحة لمواجهة تفشي كورونا في بغداد Popular mobilization units launches the process of curbing the pandemic to confront Corona outbreak in Baghdad	Thankful	Official
8	عند الشدائد تظهر الأخلاق على حقيقتها In adversity, morals appear for what they are	Empathetic	None

Furthermore, to benchmark the performance of our proposed approach, we compared the obtained results with another work using the publicly available SenWave dataset. Table 5 shows a comparison between two of our models with emojis replacement and DWLF with the other published work. As shown in the table, the proposed models outperformed the work by Yang et al. [9], which depends on the AraBERT model for sentiment analysis.

Table 5. Comparison with published work.

Model	Multi-Acc.	Jac.Acc.	F1-Micro	F1-Macro	LRAP	H-Loss
Yang et al. [9]	0.905	0.589	0.630	0.520	0.661	0.111
SenWav+AraBERT	0.929	0.622	0.658	0.536	0.664	0.071
SenWave+MARBERT	0.931	0.638	0.665	0.530	0.680	0.069
SenAIT+AraBERT	0.936	0.651	0.694	0.632	0.691	0.064
SenAIT+MARBERT	0.932	0.636	0.725	0.704	0.678	0.068

5. Conclusions

In this paper, we highlighted the lack of research on sentiment analysis for the Arabic language, especially at the finer level. This study investigated to what extent accurate deep learning models can assist in understanding society's behavior during the COVID-19 pandemic. This paper proposed a multi-label emotion classifier by employing two BERT-based transformers, AraBERT and MARBERT, with emojis replacement. We also proposed a DWLF technique to give more weight in the loss function for the samples of the minority classes. In addition, we have created a new dataset, SenAIT, by merging the common emotions of SenWave and AIT datasets. A series of experiments have been conducted

using several preprocessing steps with different word embeddings and architectures. The proposed models achieved state-of-the-art results on a benchmarked dataset, SenWave. The best performance was achieved with fine-tuning MARBERT model with emojis replacement and DWLF. The obtained results show the importance of emojis description for sentiment analysis and how DWLF can boost the performance of these systems. However, the proposed model needs to be tested on other datasets to evaluate its generalization. In addition, the proposed model has not been evaluated on other languages, which need to be investigated in the future.

As for future work, other models and datasets will be evaluated. In addition, we will evaluate the model on other languages. Moreover, the impact of morphological analysis of Arabic text on the sentiment analysis needs to be investigated further.

Author Contributions: Conceptualization, N.A. and H.L.; methodology, N.A. and H.L.; writing—original draft preparation, N.A. and H.L.; writing—review and editing, N.A. and H.L.; supervision, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge the support provided by King Fahd University of Petroleum and Minerals (KFUPM) during this work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolution Neural Network
COVID-19	Coronavirus Disease
DWLF	Dynamically weighted loss function
NLP	Natural Language Processing
SVM	Support Vector Machines
VADER	Valence Aware Dictionary and Sentiment Reasoner
WHO	World Health Organization

References

- World Health Organization. Coronavirus Disease (COVID-19). Available online: [who.int/emergencies/diseases/novel-coronavirus-2019](https://www.who.int/emergencies/diseases/novel-coronavirus-2019) (accessed on 1 May 2021).
- Appel, G.; Grewal, L.; Hadi, R.; Stephen, A.T. The future of social media in marketing. *J. Acad. Mark. Sci.* **2020**, *48*, 79–95. [[CrossRef](#)] [[PubMed](#)]
- Liu, B. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. [[CrossRef](#)]
- Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive attention networks for aspect-level sentiment classification. *arXiv* **2017**, arXiv:1709.00893.
- Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of Emotion*; Academic Press: New York, NY, USA, 1980; pp. 3–33.
- Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [[CrossRef](#)]
- Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [[CrossRef](#)]
- Oberländer, L.A.M.; Klinger, R. An analysis of annotated corpora for emotion classification in text. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2104–2119.
- Yang, Q.; Alamro, H.; Albaradei, S.; Salhi, A.; Lv, X.; Ma, C.; Alshehri, M.; Jaber, I.; Tifratene, F.; Wang, W.; et al. Senwave: Monitoring the global sentiments under the Covid-19 pandemic. *arXiv* **2020**, arXiv:2006.10842.
- Mohammad, S.M. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text editor. *Emot. Meas.* **2016**, 201–237. [[CrossRef](#)]
- Sánchez-Rada, J.F.; Iglesias, C.A. Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Inf. Fusion* **2019**, *52*, 344–356. [[CrossRef](#)]

12. Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv* **2002**, arXiv:cs/0212032.
13. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [[CrossRef](#)]
14. Zhao, J.; Liu, K.; Xu, L. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*; MIT Press One Rogers Street: Cambridge, MA, USA, 2016.
15. Medford, R.J.; Saleh, S.N.; Sumarsono, A.; Perl, T.M.; Lehmann, C.U. An “Infodemic”: Leveraging high-volume twitter data to understand early public sentiment for the Coronavirus disease 2019 outbreak. *Open Forum Infect. Dis.* **2020**, *7*, ofaa258. [[CrossRef](#)]
16. Sharma, K.; Seo, S.; Meng, C.; Rambhatla, S.; Liu, Y. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv* **2020**, arXiv:2003.12309.
17. Zhou, J.; Yang, S.; Xiao, C.; Chen, F. Examination of Community Sentiment Dynamics Due To Covid-19 Pandemic: A Case Study From Australia. *arXiv* **2020**, arXiv:2006.12185.
18. Hutto, C.J.; Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, MI, USA, 1–4 June 2014.
19. Chakraborty, K.; Bhatia, S.; Bhattacharyya, S.; Platos, J.; Bag, R.; Hassanien, A.E. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Appl. Soft Comput. J.* **2020**, *97*, 106754. [[CrossRef](#)]
20. Samuel, J.; Ali, G.G.N.; Rahman, M.M.; Esawi, E.; Samuel, Y. COVID-19 public sentiment insights and machine learning for tweets classification. *Information* **2020**, *11*, 314. [[CrossRef](#)]
21. Wrycza, S.; Maślankowski, J. Social Media Users’ Opinions on Remote Work during the COVID-19 Pandemic. Thematic and Sentiment Analysis. *Inf. Syst. Manag.* **2020**, *37*, 288–297. [[CrossRef](#)]
22. Sattar, N.S.; Arifuzzaman, S. Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the usa. *Appl. Sci.* **2021**, *11*, 6128. [[CrossRef](#)]
23. Deng, L.; Yu, D. Deep learning: Methods and applications. *Found. Trends Signal Process.* **2013**, *7*, 197–387. [[CrossRef](#)]
24. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, 1–25. [[CrossRef](#)]
25. Nassif, A.B.; Elnagar, A.; Shahin, I.; Henno, S. Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities. *Appl. Soft Comput.* **2020**, *98*, 106836. [[CrossRef](#)]
26. Habimana, O.; Li, Y.; Li, R.; Gu, X.; Yu, G. Sentiment analysis using deep learning approaches: An overview. *Sci. China Inf. Sci.* **2020**, *63*, 1–36. [[CrossRef](#)]
27. Yadav, A.; Vishwakarma, D.K. Sentiment analysis using deep learning architectures: A review. *Artif. Intell. Rev.* **2020**, *53*, 4335–4385. [[CrossRef](#)]
28. Selvapriya, M.; Priscilla, G.M. Integrated feature selection (IFS) algorithm and enhanced weight based convolutional neural network (EWCNN) for social emotion classification. *Mater. Today Proc.* **2020**, in press. [[CrossRef](#)]
29. Al Sallab, A.; Hajj, H.; Badaro, G.; Baly, R.; El Hajj, W.; Bashir Shaban, K. Deep Learning Models for Sentiment Analysis in Arabic. In Proceedings of the Second Workshop on Arabic Natural Language Processing, Beijing, China, 30 July 2015; pp. 9–17. [[CrossRef](#)]
30. Alhuzali, H.; Abdul-Mageed, M.; Ungar, L. Enabling deep learning of emotion with first-person seed expressions. In Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 25–35.
31. Senarath, Y.; Thayasivam, U. Exploring Deep Neural Networks and Transfer Learning for Analyzing Emotions in Tweets. *arXiv* **2020**, arXiv:2012.06025.
32. Abdullah, M.; Hadzikadicy, M.; Shaikhz, S. SEDAT: Sentiment and Emotion Detection in Arabic Text Using CNN-LSTM Deep Learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 835–840. [[CrossRef](#)]
33. Chen, T.; Xu, R.; He, Y.; Wang, X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst. Appl.* **2017**, *72*, 221–230. [[CrossRef](#)]
34. Abdul-Mageed, M.; Zhang, C.; Hashemi, A.; Moatez, E.; Nagoudi, B. AraNet: A Deep Learning Toolkit for Arabic Social Media. *arXiv* **2020**, arXiv:1912.13072v2.
35. Imran, A.S.; Daudpota, S.M.; Kastrati, Z.; Batra, R. Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets. *IEEE Access* **2020**, *8*, 181074–181090. [[CrossRef](#)]
36. Go, A.; Bhayani, R.; Huang, L. Twitter sentiment classification using distant supervision. *CS224N Proj. Rep. Stanf.* **2009**, *1*, 2009.
37. Mohammad, S.M.; Bravo-Marquez, F. WASSA-2017 shared task on emotion intensity. *arXiv* **2017**, arXiv:1708.03700.
38. Pran, M.S.A.; Bhuiyan, M.R.; Hossain, S.A.; Abujar, S. Analysis of Bangladeshi People’s Emotion during Covid-19 in Social Media Using Deep Learning. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020, Kharagpur, India, 1–3 July 2020. [[CrossRef](#)]
39. Wang, T.; Lu, K.; Chow, K.P.; Zhu, Q. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. *IEEE Access* **2020**, *8*, 138162–138169. [[CrossRef](#)]
40. Luo, Y.; Xu, X. Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. *Int. J. Hosp. Manag.* **2021**, *94*, 102849. [[CrossRef](#)]

41. Kabir, M.Y.; Madria, S. EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets. *Online Soc. Netw. Media* **2021**, *23*, 100135. [[CrossRef](#)] [[PubMed](#)]
42. Aljameel, S.S.; Alabbad, D.A.; Alzahrani, N.A.; Alqarni, S.M.; Alamoudi, F.A.; Babili, L.M.; Aljaafary, S.K.; Alshamrani, F.M. A Sentiment Analysis Approach to Predict an Individual's Awareness of the Precautionary Procedures to Prevent COVID-19 Outbreaks in Saudi Arabia. *Int. J. Environ. Res. Public Health* **2020**, *18*, 218. [[CrossRef](#)] [[PubMed](#)]
43. Addawood, A.; Alsuwailam, A.; Alohal, A.; Alajaji, D.; Alsuhaibani, J.; Aljabli, F.; Alturki, M. Tracking and Understanding Public Reaction During COVID-19: Saudi Arabia As A Use Case, 2020. Available online: aclanthology.org/2020.nlp-covid19-2.24v2.pdf (accessed on 9 November 2021).
44. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. *arXiv* **2020**, arXiv:2003.00104.
45. Mukherjee, R.; Poddar, S.; Naik, A.; Dasgupta, S. How have we reacted to the covid-19 pandemic? Analyzing changing indian emotions through the lens of twitter. *arXiv* **2020**, arXiv:2008.09035.
46. Hegazi, M.O.; Al-Dossari, Y.; Al-Yahy, A.; Al-Sumari, A.; Hilal, A. Preprocessing Arabic text on social media. *Heliyon* **2021**, *7*, e06191. [[CrossRef](#)]
47. Oudah, M.; Almahairi, A.; Habash, N. The impact of preprocessing on Arabic-English statistical and neural machine translation. *arXiv* **2019**, arXiv:1906.11751.
48. Duwairi, R.; El-Orfali, M. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *J. Inf. Sci.* **2014**, *40*, 501–513. [[CrossRef](#)]
49. Ghallab, A.; Mohsen, A.; Ali, Y. Arabic sentiment analysis: A systematic literature review. *Appl. Comput. Intell. Soft Comput.* **2020**, *2020*, 7403128. [[CrossRef](#)]
50. Qiao, Y.; Xiong, C.; Liu, Z.; Liu, Z. Understanding the behaviors of bert in ranking. *arXiv* **2019**, arXiv:1904.07531.
51. Deho, B.O.; Agangiba, A.W.; Aryeh, L.F.; Ansah, A.J. Sentiment analysis with word embedding. In Proceedings of the 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST), Accra, Ghana, 22–24 August 2018; pp. 1–4.
52. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2019**, arXiv:1810.04805.
53. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
54. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155. [[CrossRef](#)]
55. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems. 2013. Available online: <https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf> (accessed on 9 November 2021).
56. Soliman, A.B.; Eissa, K.; El-Beltagy, S.R. AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. *Procedia Comput. Sci.* **2017**, *117*, 256–265. [[CrossRef](#)]
57. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *arXiv* **2020**, arXiv:2101.01785.
58. Lawrence, S.; Giles, C.L.; Tsoi, A.C.; Back, A.D. Face recognition: A convolutional neural-network approach. *IEEE Trans. Neural Netw.* **1997**, *8*, 98–113. [[CrossRef](#)]
59. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
60. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. *J. Mach. Learn. Res.* **2011**, *15*, 315–323.
61. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
62. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
63. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. Semeval-2018 task 1: Affect in tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 1–17.
64. Rengasamy, D.; Jafari, M.; Rothwell, B.; Chen, X.; Figueredo, G.P. Deep Learning with Dynamically Weighted Loss Function for Sensor-Based Prognostics and Health Management. *Sensors* **2020**, *20*, 723. [[CrossRef](#)]
65. Wu, X.Z.; Zhou, Z.H. A unified view of multi-label performance measures. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.