

## Attention-Based Bi-LSTM Model for Arabic Depression Classification

Abdulqader M. Almars\*

College of Computer Science and Engineering, Taibah University, Yanbu, Saudi Arabia

\*Corresponding Author: Abdulqader M. Almars. Email: amars@taibahu.edu.sa

Received: 12 August 2021; Accepted: 14 October 2021

**Abstract:** Depression is a common mental health issue that affects a large percentage of people all around the world. Usually, people who suffer from this mood disorder have issues such as low concentration, dementia, mood swings, and even suicide. A social media platform like Twitter allows people to communicate as well as share photos and videos that reflect their moods. Therefore, the analysis of social media content provides insight into individual moods, including depression. Several studies have been conducted on depression detection in English and less in Arabic. The detection of depression from Arabic social media lags behind due the complexity of Arabic language and the lack of resources and techniques available. In this study, we performed a depression analysis on Arabic social media content to understand the feelings of the users. A bidirectional long short-term memory (Bi-LSTM) with an attention mechanism is presented to learn important hidden features for depression detection successfully. The proposed deep learning model combines an attention mechanism with a Bi-LSTM to simultaneously focus on discriminative features and learn significant word weights that contribute highly to depression detection. In order to evaluate our model, we collected a Twitter dataset of approximately 6000 tweets. The data labelling was done by manually classifying tweets as depressed or not depressed. Experimental results showed that the proposed model outperformed state-of-the-art machine learning models in detecting depression. The attention-based Bi-LSTM model achieved 0.83% accuracy on the depression detection task.

**Keywords:** Depression detection; social media; deep learning; Bi-LSTM; attention model

### 1 Introduction

Depression is one of the increasingly serious public health issues in modern society. Depression can be defined as a mental illness that causes sadness and loss of interest. In its worst case, depression can sometimes result in suicide [1,2]. According to World Health Organization research [3], the number of people affected by depression globally exceeds 350 million. Moreover, the risk of suicide is more than 25 times greater in people who suffer from depression than in the population without these illnesses [2]. The risk factors for depression can be divided into three general categories. First, the



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

most common factors that lead to depression are lose of friends or family, financial problems, social isolation, emotional pain, and a loss of hope. These factors can make people feel depressed, sometimes leading to suicidal behaviour [4,5]. Second, traumatic experiences, such as cyberbullying, sexual and physical abuse, and academic failure, often lead to depression [6]. Third, physical illnesses with little or no hope of cures, such as cancer, brain injury, diabetes, etc., have been found in many studies to lead to depression [7].

Social networks such as Twitter and Weibo have grown in popularity for public discussions on social matters. Twitter is a free broadcast channel that offers registered users the ability to discuss and communicate with others via 140-character texts. According to a recent Twitter report [8], a total of 1.3 billion accounts have been created, and there are approximately more than 320 million active users monthly and 140 million users who post over 450 million tweets each day. The popularity of these platforms provides individuals with a free space to talk about their feelings and moods. Fig. 1 shows an example taken from Twitter of users who are depressed and are expressing negative intentions, including discussion of suicide, because of cyberbullying and COVID-19.



**Figure 1:** Examples of users expressing depression using Twitter

Several machine-learning techniques, including naive Bayes (NB), support vector machine (SVM) and logistic regression (LR) have been employed in depression identification tasks [9–11]. Deep-learning techniques, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs) and long short-term memory (LSTM) can also be used for depression detection problems [12–14]. Deep learning has already had a significant influence on the field of text categorization when compared to classic machine-learning algorithms. Many studies have shown that deep neural networks outperform more traditional machine learning in terms of accuracy and performance [15,16].

Social media platforms provide people with a free space to share their emotions and thoughts in written form. This highlights the potential of social media as a rich and consistent source for analyzing posts in Twitter that are particularly relevant for health disorders analysis. Most current studies that focus on analyzing depression in social media do so exclusively in the English language and ignore other languages, such as Arabic. Arabic is the sixth most widely spoken language on the

earth. However, depression analysis in the Arabic language is still to be done. Compared with English content, the Arabic language presents more complex challenges due to the use of unstructured non-grammatical slang. In addition, depression is a global issue, and the Arab community is also affected by it; however, there have been few studies on the detection of Arabic depression.

Hence, in this study, we were interested in filling this gap by introducing a robust model for capturing behavioural health disorders from Arabic content. To achieve this, we present an effective deep learning algorithm to automatically detect depression in social media by combining Bi-LSTM and attention mechanism. The underlying intuition is to learn the discriminative features for depression detection by adding an attention mechanism to Bi-LSTM. Moreover, employing an attention mechanism can help select informative words that are more likely to contribute toward depression detection in social media posts. In this work, we mainly contribute the following.

- We introduce a novel deep learning model that learns to perform depression detection automatically from Arabic posts.
- The attention mechanism is embedded into Bi-LSTM to enable informative feature extraction from texts. As a result, the deep learning classifier can detect depression better in Arabic content.
- Our experiment results indicate that our model outperformed the other models and enhanced the accuracy of the predictions.

The remainder of the paper includes the following sections: Related work is presented in Section 2, the proposed methodology is introduced in Section 3, experimental results are presented in Section 4, and then we summarize and propose possible future directions in Section 5.

## 2 Related Work

We discuss, In this section, the literature work on depression detection and the attention mechanism briefly.

### 2.1 Depression Detection

The task of predicting depression can be considered a classification problem. Several previous research studies have been concerned with identifying depression. One popular method is to ask people to fill questionnaires or participate in interviews on social media [17,18]. For example, Park et al. [18] analyzed the mental disorder of people by exploring the use of languages of users on social media. However, these types of techniques are costly and time-consuming. In the past decade, social media platforms have been widely used for mental disorder analysis. The multimodal depressive dictionary learning model (MDL) for mental health disorder detection on Twitter was introduced by Shen et al. [19]. They first applied rule-based heuristic methods to build a well-labelled depression dataset. Then, MDL is used to learn the sparse user representation from each feature group as a single modality. The experimental results demonstrate that the model outperformed the baseline methods in terms of F1-Measure.

Kim et al. [20] used XGBoost and convolutional neural networks (CNN) to analyze texts from Reddit to determine whether a user has depression, anxiety, or a personality disorder. Mustafa et al. [21] utilize the fourteen psychological attributes in Linguistic Inquiry and Word Count (LIWC) to analyze emotions and identify depression. Based on the weights assigned by LIWC, a machine learning classifier was trained to classify users into three categories of depression. This model was evaluated using a dataset including 179 depressive individuals. A model of public reactions on Twitter was developed by Noor et al. [22] to identify the people's emotions related to the Covid-19 pandemic.

SPM methods are employed to discover frequent words/patterns within tweets, along with their relationships.

More recently, a hybrid model has been introduced that analyzes user's textual posts [23]. In a hybrid model, Bidirectional Long Short-Term Memory (Bi-LSTM) with various word representation methods and were employed to detect depression, which gave good results. Orabi et al. [24] built three models that use CNN, and the last one uses RNN on the top of the word-embeddings to detect depression of Twitter users. However, the number of tweets used to evaluate the model is relatively small. An early risk detection error (ERDE) metric is proposed by Losada et al. [25]. ERDE technology can be applied in a variety of applications, including depression identification. The system provides scores to users who are classified as depressed based on observations from the data. The main drawbacks of this model are that it includes four different meta-parameters that need to be defined before its use, and it needs to observe enough data before making a prediction. Sadeque et al. [26] addressed the drawback of ERDE by proposing a new metric called latency weighted F1 for analyzing social media posts and identifying whether a user was depressed. Other successful models in the text classification task have been proposed that apply different n-gram weighting techniques. Mac Kim et al. [27] utilized term frequency-inverse document frequency (TF-IDF) and post-level embedding for feature representations. The system employed TF-IDF weightings of unigrams to generate feature representations of the labelled dataset, whereas the embeddings were obtained using sent2vec [28]. Malmasi et al. [29] utilized lexical features to represent words using different n-grams and syntactic features for deeper linguistic analysis, using the Stanford NLP library. However, these studies applied simple classification models, such as linear SVM classifiers and the SGDClassifier.

Few studies have been done to detect depression in Arabic content. Almouzi et al. [30] applied several machine learning methods (Random Forest, Naïve Bayes, AdaBoostM1, and Liblinear) to analyze Arabic tweets and classify them as depressed or not. The results illustrate that the Liblinear algorithm archives the highest accuracy compared with other methods. However, the dataset used in this paper was unbalanced since the dataset contains 27 depressed users and 62 non-depressed users. Recently, Al-Laith et al. [31] introduced a hybrid framework to monitor people's mental and emotional issues regarding the Covid-19 pandemic. First, the rule-based method was utilized to annotate tweets using an Arabic emotion lexicon. The tweets are then classified into six categories (fear, joy, sorrow, disgust, anger) and two types (symptom and non-symptom tweets) using long-short-term memory (LSTM).

## 2.2 Attention Mechanism

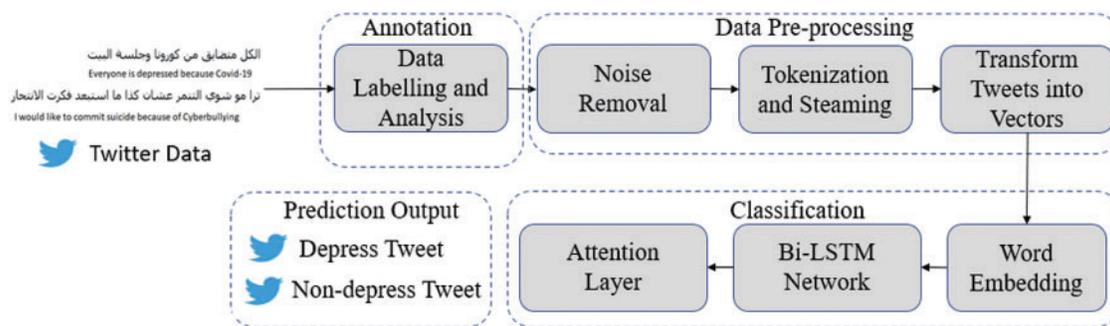
An attention mechanism was developed by Bahdanau et al. [32] to boost the performance of a neural network's encoder-decoder architecture. The core idea of attention is that it allows the model to put more attention to certain words when processing the data. In other words, the model uses only the relevant information of the input data to predict the output words. Vinyals et al. [33] have extended and improved the previous model by introducing a mechanism that computes attention vectors to reflect how much weight should be assigned to different words to enhance accuracy and performance on large-scale translations. The attention mechanism has been utilized to increase prediction accuracy in classification tasks [34–36]. Liu et al. [35] presented a deep-learning model based on attention to handling the issue of high dimensionality and sparsity in text data. Six sentiment datasets and a question dataset are used for experimental verifications. Ran et al. [36] introduced a Short-Term Long Memory (LSTM) with an attention mechanism to detect mental health disorders from interview transcriptions. The evaluation results show that the model performed the best in terms of performance compared with LSTM and SVM. Sharma et al. [37] added a SoftMax function to the hidden state

of LSTM to capture more valuable information for visual recognition. However, all mentioned papers applied an attention mechanism for videos and speech depression detection and translation tasks [33,38–40].

In conclusion, although many applications for text classification have successfully utilized the attention mechanism. This is the first study we know that uses an attention mechanism to analyze textual data related to depression. Moreover, the focus of the current research is on English with fewer contributions in Arabic. Due to the complexity of Arabic, working with Arabic text can be challenging. To overcome this issue, we found that adding an attention mechanism to a deep-learning model can help better analyze Arabic content and detect depression in terms of performance.

### 3 Proposed Model

In this section, we explain in detail the proposed model for depression prediction. Fig. 2 illustrates the flow diagram of the proposed model. In the first step, we collected and labelled the data manually. Second, we pre-process and clean the dataset and convert the word strings to a sequence of integers. Third, we then pre-trained our model to build the vector representations of the words. Fourth, we explain the attention mechanism that can be added to the Bi-LSTM to focus on particular features and enhance the accuracy of predictions for the depressed and not depressed binary classification. In the following subsections, we explained the steps in detail.



**Figure 2:** The flow diagram of the proposed model

#### 3.1 Data Collection

A dataset was collected using the Twitter Developer Application Programming Interface (API). Unfortunately, due to a Twitter policy restriction, we only collected approximately 6000 tweets related to depression written in the Arabic language. The dataset was collected between the 17<sup>th</sup> of February and the 8<sup>th</sup> of April. Tab. 1 shows the statistical details of the dataset. To balance the dataset, we gathered a list of tweets related to depression using the following hashtags: (1) *مكتئاب* (depress), (2) *زعلان* (upsets), and (3) *حزين* (sad). In addition, to balance the ratio of depressed and not depressed, we used the following hashtags: (1) *مبسوط, سعيد* (happy) and (2) *متفائل* (optimistic). After collecting the dataset, extensive cleaning and preprocessing were done to remove noise and outliers. Each tweet was then labelled as one for depressed or zero for not depressed. In the following subsections, we describe the annotation step and the preprocessing step in detail.



all the words in the text, and each word has a unique index. By completing this step, you can create a dictionary mapping word to an index, which will be needed for the word embedding step. To make the data uniformly distributed, we also apply the same scale on the data by padding with zeros. In other words, each sentence with a sequence length shorter than the mean length is padded with zeros, while longer sentences are reduced to the mean sequence length.

### 3.4 Word Embedding

Word embedding is a popular approach to extracting word associations and relations from a large text corpus in natural language processing. These words are visualized as vectors of real numbers. Such a model can help to find synonymous words or suggest a list of words for a partial sentence. Various models have been introduced for learning word embeddings from raw text, such as GloVe [42] and the Word2Vec model [43]. Word2Vec has two model structures for learning feature representations, the continuous bag-of-words (CBOW) and the skip-gram. Pre-trained word embedding is a popular techniques used to improve the accuracy of neural network models. In Arabic content domains, there are a few open-source pre-trained models. In our experiment, we use the pre-trained AraVec [44] to learn distributed word representations. This model uses Word2vec to learn word vectors from sources such as the Twitter platform, Wikipedia articles, and other Internet content. Moreover, AraVec contains a rich vocabulary that can help our model capture more semantic signals from the pre-trained embeddings. The total number of vectors used to build AraVec is approximately 3.300.000.000.

### 3.5 Bidirectional Long Short-Term Memory (Bi-LSTM)

LSTM is a deep-learning model that has been developed and implemented successfully for a variety of tasks such as text categorization, picture processing, and natural language processing. It has shown great performance when compared with other deep-learning methods [45]. Unidirectional LSTM only keeps and remembers contextual information of the past, and it neglects the future contextual data [45]. In contrast, bidirectional LSTM (Bi-LSTM) [46] addresses this issue by using two independent LSTMs (forward and backward). The forward LSTM processes data in the forward direction (from the past to the future), whereas the backward LSTM processes data in the opposite direction (from the future to the past). Fig. 3 illustrates the proposed attention-based deep learning model. In our proposed model, we employ the Bi-LSTM network to learn the discriminative word representations for depression detection. The structure of LSTM consists of (1) forget gate  $f_t$  (2) input gate  $i_t$  (3) output gate  $o_t$  and (4) cell memory state  $c_t$ . Given the input features  $x = \{x_1, x_2, x_3, \dots, x_t\}$ , the hidden state can be computed as follows:

$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

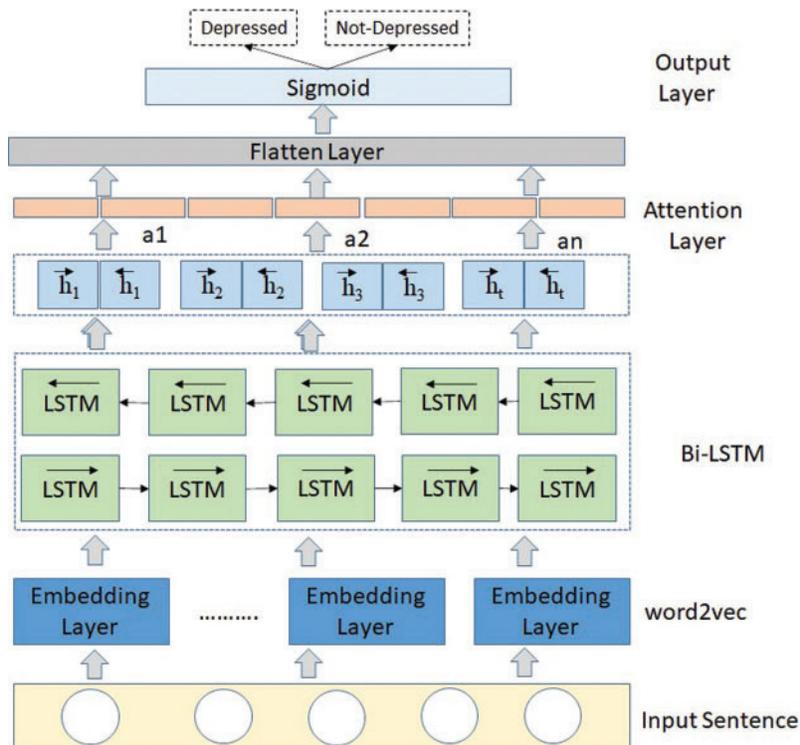
$$i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$o_t = \sigma (W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh( (W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$h_t = o_t \odot \tanh(c_t) \quad (5)$$

where  $\alpha$  is a sigmoid function;  $W_f$ ,  $W_i$ ,  $W_o$  are the weighted matrix and  $b_f$ ,  $b_i$ ,  $b_o$  are the biases (hyperparameters) of the LSTM, which will be learned during the training process; and  $x_t$  is the inputs of the LSTM, and  $h_t$  is the vector of the hidden layer. The forward hidden state  $\vec{h} = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_t\}$  is computed using the forward LSTM, while the backward hidden  $\overleftarrow{h} = \{\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_t\}$  is computed using the backward LSTM. The output of hidden state of the model is then produced by concatenating  $h_t = \{\vec{h}_t, \overleftarrow{h}_t\}$ . It is important to mention that our problem is a binary classification task. The sigmoid function is used here for the binary depression classification.



**Figure 3:** Attention-based Bi-LSTM model

### 3.6 Attention Layer

When dealing with text classification tasks, it is important to understand the relations between words to correctly label the text to either positive “depress” or negative “not-depress”. However, not all words in a sentence are equally significant in the text classification task. To overcome this issue, several types of attention mechanisms have been proposed and applied to various tasks. The idea behind the attention mechanism is that it places greater emphasis on important content by assigning higher weights to specific keywords. This mechanism has shown an excellent performance in several text classification applications.

In our model, we implemented an attention mechanism similar to the implementation of Bahdanau et al. [32] for analyzing Arabic tweets. By adding an attention layer, we allow our model to decide which words should receive “attention”. In the same sentence, some words are highly related to depression while others are irrelevant. For example, words related to depression such as “sad, depressed, unhappy . . . etc.” can receive more weight and enables the classifier to perform better in

depression detection. Hence, the attention layer can help to focus on specific information. To achieve this, our model first receives the hidden  $h_t$  vectors at each time step from the Bi-LSTM model; the model then learns the hidden representation  $m_t$  and context vector  $m_w$ . The context vector  $m_w$  here represents the weights (importance) of different words over the word sequences. The attention-based model uses a Softmax function to calculate the weighted mean of the state  $h_t$ . In other words, we assign a single value to each vector of representation, which we call an attention weight. The formula is described as follows:

$$m_t = \tanh(W_w h_t + b_w), \quad (6)$$

$$a_t = \text{softmax} \frac{\exp(m_t^T m_w)}{\sum_t \exp(m_t^T m_w)}, \quad (7)$$

$$s = \sum_t a_t h_t. \quad (8)$$

where  $\alpha_t$  is the attention weight vector, and the weighted hidden representations are  $s$ . The outputs are then passed to a fully connected Sigmoid layer which is then used to determine whether the input text is depressed or not.

## 4 Experiments

In this section, we discuss the evaluation of the performance and accuracy of our suggested model using a dataset collected from Twitter. We first describe the experimental settings. Then, our model's performance was calculated by comparing it to other approaches using precision, recall, F-measure, and accuracy.

### 4.1 Experimental Settings

In our model, there are a number of parameters to tune. Several experiments were conducted to determine the model's best parameters. The performance of the proposed model in different hyperparameters is illustrated in [Tab. 3](#). The following optimal configurations were used. The dropout rate was 0.2; the hidden layer size was 64, and the number of epochs was 20. We chose the Adam optimizer and set the learning rate at 0.001. In our model, the loss function (binary cross-entropy) was used. Early stopping was employed to assess the model's performance on the dataset. The total number of parameters of the model are shown in [Tab. 4](#). To set up our model, we first divide our dataset into 70% for training and 30% for testing. As we are dealing with short texts, we fixed the word embedding size to 50 with a length of 30.

### 4.2 Evaluation Metrics

The performance of our model was measured using four well-known metrics: precision, recall, F-measure, and accuracy. The equations are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FN} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FP} \quad (10)$$

$$F - Measure = 2 \cdot \frac{precision * recall}{precision + recall} \quad (11)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (12)$$

where (TP) is the number of genuine samples accurately labelled as positive. (FP) denotes the total number of false samples that are wrongly labelled as positive. The total number of false samples that are correctly labelled as negative is (TN), whereas the number of real samples that are wrongly labelled as negative is (FN).

**Table 3:** Performance of the proposed model in different hyperparameters

No. of layers	Alpha	Accuracy (%)
64	0.1	79
	0.01	80
	0.001	83
512	0.1	62
	0.01	78
	0.001	80
1024	0.1	56
	0.01	77

**Table 4:** The total number of parameters of the model

Layer (type)	Output shape	Param #
Input	[(None, 30)]	0
Embedding	(None, 30, 50)	751,450
Bi-LSTM	(None, 30, 64)	7920
Attention	(None, 30, 64)	1985
Flatten	(None, 750)	0
Dense	(None, 1)	715
<b>Total params</b>		762,106
<b>Trainable params</b>		762,106

### 4.3 Experimental Results

Our goal in this section is to evaluate the proposed depression detection model in Section 4 against the following state-of-the-art baselines that utilize textual content for depression detection:

- Machine learning models like naïve bayes, support vector machines (SVM) and Random Forests have been used recently to detect mental health conditions in user posts [9–11,47].
- XGBoost [48] is a machine learning technique that has been used recently to analyze texts from social media in an attempt to identify whether a user has a specific mental health disorder.

- The LSTM model has been widely explored in previous depression detection studies [31,49].
- CNN: Several previous studies have used CNN to detect depression automatically from Twitter [24]. We train two CNN models for baselines: CNNWithMax and MultiChannelCNN.
- Bi-LSTM model that was used in a recent study for depression classification of text [23,24].

Fig. 4 shows the most frequent words portrayed in each category (positive is “depressed”; negative is “not depressed”), as visualized using the WordCloud module. Positive was defined in terms of the words that were the target of this study not the emotional valence of the words. As we mentioned earlier, our proposed model attempts to identify depression from tweets. Tweets are classified based on their language, emotional level, suicidal tendencies, and factors associated with depression. Users are categorized as being depressed if they use specific words related to depression, otherwise non-depressed (normal). Tab. 5 displays the prediction results for our model compared with others. As illustrated in the table, in terms of F-measure, naive Byes performed the lowest when compared with the other models. It had an F-measure of 0.77 in the negative tweets (not-depressed) and 0.74 in the positive ones (depressed). For the SVM and MultiChannelCNN, the results show that they performed better than the Naïve Byes in depression detection with F-measures of 0.82 and 0.77 for the negative and positive tweets, respectively. Random forest, XGBoost, CNNWithMax, LSTM and Bi-LSTM achieved better results, indicating the ability of those models to detect depressions from social media. However, the model’s performance was enhanced by combining the Bi-LSTM with an attention mechanism in precision, recall and F-measure. When compared to the baseline models, the attention-based LSTM model generated the highest F-measure for positive and negative tweets. As Fig. 5 shows, the proposed model obtains the greatest accuracy of 83% compared with the other models.

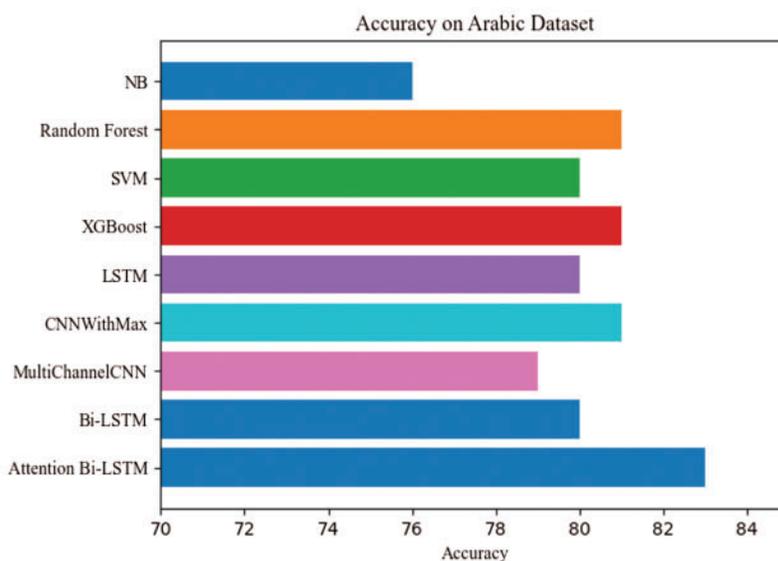


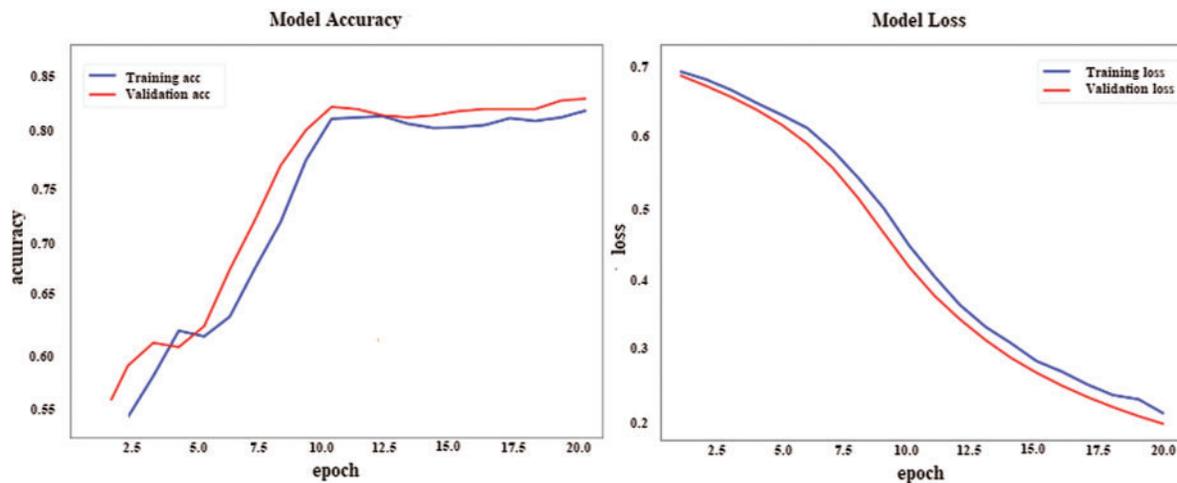
Figure 4: WordClouds for most used positive (“depressed”) and negative (“not depressed”) words

Furthermore, Fig. 6 shows the accuracies of our model over different epochs. The accuracy of the proposed model confirmed our assumption that embedding an attention layer with a Bi-LSTM would achieve better results when dealing with Arabic content. To sum up, two main observations can be concluded from the outcomes of our model. First, the deep-learning model outperformed classic machine learning models such as naive Bayes and SVM. Second, adding an attention mechanism to the Bi-LSTM can help it focus on discriminative words and enhance the model’s performance, which indicates that the model can successfully recognize depressed and non-depressed tweets.

**Table 5:** Performance of our proposed model compared with others

Method	Label	Precision	Recall	F-Measure
Naive Bayes	Negative	0.85	0.71	0.77
	Positive	0.67	0.82	0.74
SVM	Negative	0.84	0.81	0.82
	Positive	0.75	0.78	0.77
Random forest	Negative	0.85	0.81	0.83
	Positive	0.78	0.81	0.79
XGBoost	Negative	0.84	0.82	0.83
	Positive	0.78	0.82	0.80
LSTM	Negative	0.85	0.80	0.83
	Positive	0.75	0.81	0.78
CNNWithMax (CBOW)	Negative	0.80	0.79	0.80
	Positive	0.77	0.78	0.78
MultiChannelCNN (CBOW)	Negative	0.77	0.80	0.79
	Positive	0.74	0.78	0.76
Bi-LSTM	Negative	0.84	0.81	0.82
	Positive	0.79	0.80	0.80
Attention Bi-LSTM	Negative	<b>0.87</b>	<b>0.84</b>	<b>0.85</b>
	Positive	<b>0.78</b>	<b>0.83</b>	<b>0.81</b>

**Figure 5:** The accuracy of the proposed model compared with other



**Figure 6:** Accuracy and loss of the proposed model in Arabic dataset

## 5 Conclusion

In this study, we addressed the issue of identifying depression on Twitter. Tackling this problem is essential for understanding the reasons for depression expressed in social media. In this paper, we propose an attention-based Bi-LSTM to analyze and detect depression in Arabic content. The proposed model can successfully analyze and classify tweets as depressed or not depressed by adding an attention mechanism to a Bi-LSTM model. A WordCloud was then used to visualize the most frequently found Arabic words to express depression in our social media dataset. The effectiveness and accuracy of our model were evaluated with the Twitter dataset. The findings of our experiments indicate that our proposed model outperformed the other tested classification approaches. As future work, an exciting and possible direction would be incorporating non-textual features into our proposed model to improve depression detection in social media. Moreover, we anticipate reporting the performance of our model on a large-scale dataset.

**Funding Statement:** The author received no specific funding for this study.

**Conflicts of Interest:** The author declares that they have no conflicts of interest to report regarding the present study.

## References

- [1] U.S. Department of Health and Human Services, “*Healthy People 2010: Understanding and Improving Health*,” Government Printing Office, Washington, DC, 2000.
- [2] S. B. Guze and E. Robins, “Suicide and primary affective disorders,” *The British Journal of Psychiatry*, vol. 117, no. 539, pp. 437–438, 1970.
- [3] C. J. Murray, A. D. Lopez and W. H. Organization, “The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: Summary,” *World Health Organization*, vol. 1, pp. 44–55, 1996.
- [4] B. A. Primack, A. Shensa, C. G. Escobar-Viera, E. L. Barrett, J. E. Sidani *et al.*, “Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among us young adults,” *Computers in Human Behavior*, vol. 69, pp. 1–9, 2017.

- [5] L. Y. Lin, J. E. Sidani, A. Shensa, A. Radovic, E. Miller *et al.*, “Association between social media use and depression among us young adults,” *Depression and Anxiety*, vol. 33, no. 4, pp. 323–331, 2016.
- [6] E. M. Selkie, R. Kota, Y.-F. Chan and M. Moreno, “Cyberbullying, depression, and problem alcohol use in female college students: A multisite study,” *Cyberpsychology, Behavior, and Social Networking*, vol. 18, no. 2, pp. 79–86, 2015.
- [7] G. Coppersmith, C. Harman and M. Dredze, “Measuring post-traumatic stress disorder in twitter,” in *Proc. Int. AAAI Conf. on Web and Social Media*, Michigan, USA, vol. 8, no. 1, pp. 579–582, 2014.
- [8] Aslam, S. Twitter by the Numbers: Stats, Demographics & Fun Facts. 2021. Available online: <https://www.omnicoreagency.com/twitterstatistics/#:~:text=Twitter%20Demographics&text=There%20ar%20262%20million%20International,users%20have%20higher%20college%20degrees> (accessed on 11 February 2021).
- [9] N. Al Asad, M. A. M. Pranto, S. Afreen and M. M. Islam, “Depression detection by analyzing social media posts of user,” in *2019 IEEE Int. Conf. on Signal Processing, Information, Communication & Systems (SPICSCON)*, Dhaka, Bangladesh, pp. 13–17, 2019.
- [10] M. Deshpande and V. Rao, “Depression detection using emotion artificial intelligence,” in *2017 Int. Conf. on Intelligent Sustainable Systems (ICISS)*, Palladam, India, pp. 858–862, 2017.
- [11] S. Alghowinem, R. Goecke, M. Wagner, G. Parker and M. Breakspear, “Eye movement analysis for depression detection,” in *2013 IEEE Int. Conf. on Image Processing*, Melbourne, Australia, pp. 4220–4224, 2013.
- [12] Y. Wang, M. Huang, X. Zhu and L. Zhao, “Attention-based lstm for aspect-level sentiment classification,” in *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing*, Austin, Texas, pp. 606–615, 2016.
- [13] H. Dibeklioglu, Z. Hammal and J. F. Cohn, “Dynamic multimodal measurement of depression severity using deep auto encoding,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2017.
- [14] B. Ay, O. Yildirim, M. Talo, U. B. Baloglu, G. Aydin *et al.*, “Automated depression detection using deep representation and sequence learning with eeg signals,” *Journal of Medical Systems*, vol. 43, no. 7, pp. 1–12, 2019.
- [15] A. Almars, X. Li and X. Zhao, “Modelling user attitudes using hierarchical sentiment-topic model,” *Data & Knowledge Engineering*, vol. 119, pp. 139–149, 2019.
- [16] A. Almars, X. Li, X. Zhao, I. A. Ibrahim, W. Yuan *et al.*, “Structured sentiment analysis,” in *Int. Conf. on Advanced Data Mining and Applications*, NY, USA, pp. 695–707, 2017.
- [17] O. Whooley, “Diagnostic and statistical manual of mental disorders (dsm),” *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*, vol. 5, pp. 381–384, 2014.
- [18] M. Park, C. Cha and M. Cha, “Depressive moods of users portrayed in twitter,” in *Proc. of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, Beijing, China, pp. 1–8, 2012.
- [19] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang *et al.*, “Depression detection via harvesting social media: A multimodal dictionary learning solution,” in *Proc. of the 26th Int. Joint Conf. on Artificial Intelligence (IJCAI-17)*, Melbourne, Australia, pp. 3838–3844, 2017.
- [20] J. Kim, J. Lee, E. Park and J. Han, “A deep learning model for detecting mental illness from user content on social media,” *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.
- [21] R. U. Mustafa, N. Ashraf, F. S. Ahmed, J. Ferzund, B. Shahzad *et al.*, “A multiclass depression detection in social media based on sentiment analysis,” in *Proc. of the 17th IEEE Int. Conf. on Information Technology—New Generations*, Cham, Springer, pp. 659–662, 2020.
- [22] S. Noor, Y. Guo, S. H. H. Shah, P. Fournier-Viger and M. S. Nawaz, “Analysis of public reactions to the novel coronavirus (COVID-19) outbreak on twitter,” *Kybernetes*, vol. 50, no. 5, pp. 1633–1653, 2020.
- [23] F. M. Shah, F. Ahmed, S. K. S. Joy, S. Ahmed, S. Sadek *et al.*, “Early depression detection from social network using deep learning techniques,” in *2020 IEEE Region 10 Symposium (TENSYP)*, Dhaka, Bangladesh, pp. 823–826, 2020.

- [24] A. H. Orabi, P. Buddhitha, M. H. Orabi and D. Inkpen, "Deep learning for depression detection of twitter users," in *Proc. of the 5th Workshop on Computational Linguistics and Clinical Psychology: from Keyboard to Clinic*, New Orleans, LA, pp. 88–97, 2018.
- [25] D. E. Losada, F. Crestani and J. Parapar, "Erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations," in *Int. Conf. of the Cross-Language Evaluation Forum for European Languages*, Cham, Springer, pp. 346–360, 2017.
- [26] F. Sadeque, D. Xu and S. Bethard, "Measuring the latency of depression detection in social media," in *Proc. of the 11th ACM Int. Conf. on Web Search and Data Mining*, NY, USA, pp. 495–503, 2018.
- [27] S. Mac Kim, Y. Wang, S. Wan and C. Paris, "Data61-csiro systems at the clpsych 2016 shared task," in *Proc. of the 3rd Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, pp. 128–132, 2016.
- [28] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Int. Conf. on Machine Learning*, Beijing, China, pp. 1188–1196, 2014.
- [29] S. Malmasi, M. Zampieri and M. Dras, "Predicting post severity in mental health forums," in *Proc. of the 3rd Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, pp. 133–137, 2016.
- [30] S. Almouzini and A. Alageel, "Detecting Arabic depressed users from twitter data," *Procedia Computer Science*, vol. 163, pp. 257–265, 2019.
- [31] A. Al-Laith and M. Alenezi, "Monitoring people's emotions and symptoms from Arabic tweets during the covid-19 pandemic," *Information*, vol. 12, no. 2, pp. 86, 2021.
- [32] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. of Int. Conf. on Learning Representations*, San Diego, CA, USA, pp. 777–780, 2014.
- [33] O. Vinyals, Ł Kaiser, T. Koo, S. Petrov, I. Sutskever *et al.*, "Grammar as a foreign language," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2773–2781, 2015.
- [34] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola *et al.*, "Hierarchical attention networks for document classification," in *Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, pp. 1480–1489, 2016.
- [35] G. Liu and J. Guo, "Bidirectional lstm with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.
- [36] X. Ran, Z. Shan, Y. Fang and C. Lin, "An lstm-based method with attention mechanism for travel time prediction," *Sensors*, vol. 19, no. 4, pp. 861, 2019.
- [37] Sharma, R. Kirov and R. Salakhutdinov, "Action recognition using visual attention," *ArXiv Preprint ArXiv: 1511.04119*, 2015.
- [38] L. He, J. C.-W. Chan and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, 2021.
- [39] G. I. Winata, O. P. Kampman and P. Fung, "Attention-based lstm for psychological stress detection from spoken language using distant supervision," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 6204–6208, 2018.
- [40] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang *et al.*, "Hierarchical attention transfer networks for depression assessment from speech," in *ICASSP 2020–2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 7159–7163, 2020.
- [41] L. S. Larkey, L. Ballesteros and M. E. Connell, "Light stemming for Arabic information retrieval," in *Arabic Computational Morphology*, Springer Netherlands, vol. 38, pp. 221–243, 2007.
- [42] J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543, 2014.
- [43] Y. Goldberg and O. Levy, "Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method," *ArXiv Preprint ArXiv: 1402.3722*, 2014.
- [44] A. B. Soliman, K. Eissa and S. R. El-Beltagy, "Aravec: A set of Arabic word embedding models for use in Arabic nlp," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017.

- [45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [46] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [47] F. Cacheda, D. Fernandez, F. J. Novoa and V. Carneiro, "Early detection of depression: Social network analysis and random forest techniques," *Journal of Medical Internet Research*, vol. 21, no. 6, pp. e12554, 2019.
- [48] V. Arun, V. Prajwal, M. Krishna, B. Arunkumar, S. Padma *et al.*, "A boosted machine learning approach for detection of depression," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Bangalore, India, pp. 41–47, 2018.
- [49] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Detection of suicide ideation in social media forums using deep learning," *Algorithms*, vol. 13, no. 1, pp. 7, 2020.