Research Article

Hanane Elfaik* and El Habib Nfaoui

Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text

https://doi.org/10.1515/jisys-2020-0021
Received Feb 15, 2020; accepted Aug 30, 2020

Abstract: Sentiment analysis aims to predict sentiment polarities (positive, negative or neutral) of a given piece of text. It lies at the intersection of many fields such as Natural Language Processing (NLP), Computational Linguistics, and Data Mining. Sentiments can be expressed explicitly or implicitly. Arabic Sentiment Analysis presents a challenge undertaking due to its complexity, ambiguity, various dialects, the scarcity of resources, the morphological richness of the language, the absence of contextual information, and the absence of explicit sentiment words in an implicit piece of text. Recently, deep learning has obviously shown a great success in the field of sentiment analysis and is considered as the state-of-the-art model in Arabic Sentiment Analysis. However, the state-of-the-art accuracy for Arabic sentiment analysis still needs improvements regarding contextual information and implicit sentiment expressed in different real cases. In this paper, an efficient Bidirectional LSTM Network (BiLSTM) is investigated to enhance Arabic Sentiment Analysis, by applying Forward-Backward encapsulate contextual information from Arabic feature sequences. The experimental results on six benchmark sentiment analysis datasets demonstrate that our model achieves significant improvements over the state-of-art deep learning models and the baseline traditional machine learning methods.

Keywords: Sentiment Analysis (SA), Bidirectional LSTM Network (BiLSTM), Deep Learning, Arabic language

1 Introduction

To find an automatic way to analyse, classify, and determine the attitude of a speaker in social networks is very important. Indeed, it is the most empirical way to get direct feedback or information from people [1]. For instance, in business, it allows companies to automatically gather the opinions of their customers about their products or services. In politics, it can help to infer the public orientation and reaction towards political events, which can help in decision making [2]. As a result, Sentiment analysis (SA), which aims at extracting people's opinions automatically, has gained interest in recent years in politics, social media, and business.

Sentiment Analysis refers to the use of NLP, computational linguistics, and data mining techniques to identify and retrieve certain sentiment(s) from text [3]. Its goal is to extract the sentiment conveyed in a piece of text (tweet, post, etc.) based on its content and its level of analysis (document level, sentence level, aspect level, and word level). Furthermore, the application of Sentiment Analysis in Arabic text is a timely subject. Given its importance as a language (Arabic which is recognized as the fifth most widely spoken language in the world and is considered the official or native language for 22 countries approximately more than 300 million people) [4, 5]. It has three varieties, which include classical Arabic that is found in religious and old scripts.

***Corresponding Author: Hanane Elfaik:** LISAC Laboratory, Faculty of Sciences Dhar EL Mehraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco; Email: hanane.elfaik@usmba.ac.ma
El Habib Nfaoui: LISAC Laboratory, Faculty of Sciences Dhar EL Mehraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco; Email: elhabib.nfaoui@usmba.ac.ma

Modern Standard Arabic (MSA) found in today's written scripts and mainly spoken in formal channels. The third is the colloquial or dialectal Arabic (DA), the spoken language in informal channels.

Arabic is both morphologically rich and highly ambiguous; it has complex morpho-syntactic agreement rules and a lot of irregular forms, and it has a large number of dialectal variants with no writing standards. Without proper processing and handling, learning robust general models over Arabic text can be hard to do. Furthermore, compared to English, there are less freely available resources for Arabic Sentiment analysis in terms of Sentiment lexicons and annotated corpora of sentiments. These challenges have fuelled extensive research interest in Arabic Sentiment Analysis (ASA) [6].

Despite the importance of Arabic Sentiment Analysis, state-of-the-art artificial intelligence (AI) systems rely on deep learning techniques and have achieved immense success in many domains. Deep Learning (DL), a subfield of machine learning (ML), depends on a set of algorithms in order to learn multiple levels of representation with the aim of finding a model for high-level abstractions in data. One of the main benefits of deep learning over various traditional machine learning algorithms is its capacity to execute feature engineering on its own, lies in their ability to perform semantic composition-generating a vector of representation for text units by combining their finer-grained constituents or entities-efficiently and low-dimensional space.

Recurrent neural networks (RNNs) are deep learning neural networks designed specifically to learn sequences of data and are mainly used for textual data classification. However, RNNs suffer from the vanishing gradient problem when handling long sequence of data. LSTM neural networks [7] were proposed as a solution for this problem and have proven to be efficient in many real-world issues like speech recognition [8], image captioning [9], music composition [10]. However, extracting a sentiment highly depends on the review's contextual information. Direct feedforward neural networks lack to some extent the ability to take contextual information into consideration and hence act poorly in the ASA task. Therefore, the sentiment analysis approach in this paper is the use of a Bidirectional LSTM Network (BiLSTM) with the ability of extracting the contextual information from the feature sequences by dealing with both forward and backward dependencies. Moreover, our BiLSTM allows us to look ahead by employing a forward LSTM, which processes the sequence in chronological order, and a backward LSTM, which handles the sequence in reverse order. The output is then the concatenation of the corresponding states of the forward and backward LSTM.

In our proposed approach, in order to ameliorate the state-of-art performance of the Arabic Sentiment Analysis, we have deployed a deep learning model based on BiLSTM for sentiment classification. Our contribution can be summarized as follow:

1. We investigate the benefits of Arabic pre-processing such as tokenization, Punctuations removal, Latin characters removal, Digits removal, Normalization, and Stemming.
2. Our work has used Deep Neural Network Bidirectional LSTM Network with the ability of extracting the contextual information from the feature sequences of Arabic sentences.
3. Our BiLSTM model significantly outperforms, in terms of accuracy, other deep learning models (CNN and LSTM) on the majority of the benchmark datasets.
4. It successfully achieves the highest accuracy on Arabic Sentiment Analysis classification as compared to Baseline Traditional Machine Learning methods and outperforms the state-of-the-art deep learning model's accuracy.

The remainder of this paper is organized as follows. We will first overview some related recent works on Arabic sentiment analysis approaches and methods. Then, In Section 3, we will describe our Arabic Sentiment Analysis system using deep learning model in detail. Section 4 will provide the experimental study and the obtained results. Section 5 will present the discussion. Finally, section 6 as a conclusion, an outline for future work will be given

2 Related Work

Sentiment analysis aims to classify the subjective texts in two or more categories. The most obvious ones are positive and negative. We refer to this problem as Binary SA (BSA). Some works include a third class for neutral text. We refer to this problem as ternary SA (TSA). A final option is to consider sentiment based on some ranking or rating system, such as the 5-star rating system. This is known as Multi-way SA(MWSA) [11].

The available research on Arabic Sentiment Analysis approaches can be categorized into Machine Learning, Lexicon-based, and Hybrid or combined approaches.

The **machine learning** is the most commonly used approach in sentiment analysis. [12] performed SA of tweets written in Modern Standard Arabic (MSA) and Egyptian dialects. They have collected 1000 tweets (500 positives and 500 negatives). They have used standard n-gram features and experimented with several classifiers (SVM and NB). [13] developed a sentiment analysis tool for colloquial Arabic and MSA to evaluate it in social networks. They have collected 1080 Arabic reviews from social media and news sites, and K-NN was used as a classifier. In [14], They applied SVM and NB classifiers for sentiment analysis on Arabic reviews and comments from Yahoo Maktoob website. [15] developed a supervised system for Arabic social media (SAMAR) using an SVM classifier. [16] presents an Arabic sentiment analysis tool with three classifiers SVM, K-NN, and naïve Bayes.

Another approach for Arabic sentiment analysis is the **lexicon-based approach**. It is usually implemented when the data are unlabeled. Lexicons are sentiment dictionaries with the word and its occurring sentiment or sentiment score. [17] proposed a system consists of two different parts: the first part is a free online game. The aim of this game was to build a lexicon with positive and negative words. The second part was the sentiment analysis which classified reviews according to their sentiments. [18] created a lexicon with 120,000 Arabic terms. They collected a large number of articles from the Arabic news website.

[19] developed a **hybrid approach** that utilized a lexicon-based approach as well as a machine learning-based approach. The lexicon was constructed by translating the SentiStrength English lexicon. The classification step was conducted using Maximum Entropy (ME) and K-Nearest Neighbors (KNN). [20] conducted an experiment to develop an ASA system. The authors conducted three experiments. The first experiment classified comments using the SVM machine learning-based approach only. Alternatively, the second experiment used the lexicon-based approach with the lexicon only; they constructed SSWIL in order to classify the comments. The third experiment combined the two approaches, that is, classifying comments using the SSWIL lexicon and then applying the SVM machine learning-based approach. [21] built a new Arabic lexicon by merging two MSA lexicons and two Egyptian Arabic lexicons. They used SVM as a classifier and adapted one of the state-of-the-art Semantic Orientation (SO) algorithms. [22] investigated sentiment analysis with dialectical Arabic words and MSA. The classifiers they used were NB and SVM. To convert dialectical words to MSA, the authors used a dialect lexicon that contains dialectical words with their corresponding MSA words.

Deep learning is a branch of machine learning which aims to model high-level abstractions in data. This is done using model architectures that have complex structures or those composed of multiple nonlinear transformations [23]. Only a few researchers have explored deep learning models in Arabic text.

[24] explored several deep learning models: Deep neural network (DNN): DNN applies the backpropagation to a conventional neural network with several layers; Deep belief networks (DBN): DBN pre-trains phases before feeding it into other steps; Deep autoencoder (DAE): DAE reduces dimensionality to original models; Combined DAE with DBN; and Recursive autoencoder (RAE): The RAE parses raw sentence words in the best order which then minimizes the error of creating the same sentence words in the same order. The experimentation results show that the DAE model gives a better representation of the sparse input vector. The best model was the RAE leading to an accuracy of 74%. Moreover, the RAE model's performance was better than other models by around 9%.

In [25] they proposed enhancements to the RAE model, which was proposed in [24], to adapt to challenges that arise with the Arabic text. Morphological tokenization is proposed to overcome the overfitting and morphological complexity of the Arabic sentences. The model was tested with three different datasets. The same

authors [25] participated in Sem-Eval 2017 task 4 with their RAE model and achieved an accuracy equal to 41% [26].

[27] applied the recursive neural tensor network (RNTN) model on the Arabic text. The model was trained by a sentiment treebank called (ARSENTB), which is a morphologically enriched treebank created by the authors. They used word2vec embedding using the CBOW model on the QALB corpus [28].

In [29], the authors tested their model proposed in [27] with ASTD [30] dataset of 10,006 tweets. The RNTN was trained twice, first time using lemmas, where each lemma represents a set of words that have the same meaning and differ by inflectional morphology only, and in the second time using raw words. RNTN, when trained by lemmas, showed better accuracy.

[31] performed a system to analyse opinions about health services. They gathered their dataset from Twitter hashtags and ended up with 2026 tweets. They compared two DL models, namely, DNN and CNN, with word2vec embedding. The CNN model had the best accuracy. However, in this study, the CNN model was trained on a very small dataset, and the two models did not address the negation problem. Lately, the authors proposed another model in [32] to overcome the limitation of training CNN on a small dataset. Instead, they trained a combined CNN and lexicon model on top of word2vec constructed from a large corpus acquired from multiple Arabic journals. The accuracy of their model has increased from 90% to 92%. The same authors in [33] applied sentiment analysis on a health dataset using combined deep learning algorithms (CNN-LSTM). It also explored the effectiveness of using different levels of sentiment analysis. The experimentation results show that word-level and Ch5-gram-level have shown better sentiment classification results.

Another work was proposed in [34], using the DNN algorithm. The authors have used eight layers in the model to classify Arabic tweets. The sentiment of each tweet is given by extracting the sentiment words from the tweet using a lexicon and then summing their polarities. Although the model showed good performance, it exhibited sensitivity in its performance towards different datasets, and there was not any consideration for negation.

In [35] they have examined two word embedding (CBOW and SG), using a corpus with 3.4 billion Arabic words selected from 10 billion words collected by crawling web pages. Then, to classify sentiments, a CNN-based model was trained by a previously trained word embedding.

In [36], five architectures were used, including CNN, CNN-LSTM, simple LSTM, stacked LSTM, and combined LSTM to analyse Arabic tweets. They employed dynamic and static CBOW and SG word embeddings to train the models. Experiment results showed that the combined LSTM model trained by dynamic CBOW outperformed the other models.

In [2], They have used an ensemble model, combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, to predict the sentiment of Arabic tweets. Their results show a significant increase in the performance of the models, and F1-score of 64.46% was achieved with their model.

In [37], they have proposed a hybrid incremental learning model for Arabic sentiment analysis. Their model uses two different ML classifiers and one DL classifier, which is an RNN. The input to the network is a set of 16 different weights calculated from three different lexicons that form the feature vector.

[38] addressed the aspect-based sentiment analysis for Arabic Hotels reviews. Their dataset consisted of 2291 Arabic reviews. The dataset was prepared using AraNLP [39] and MADAMIRA [40] tools, which were used to extract semantic, syntactic, and morphological features. They used the RNN approach, and the network consisted of five hidden layers.

[41] proposed a DL model to analyse sentiment expressed at the aspect level. The proposed model is based on the LSTM network, where the input to the model is the text embeddings along with the aspect embedding.

A summary of the DL Arabic sentiment analysis models that have been proposed is presented in Table 1.

Table 1: Summary of deep learning models for Arabic sentiment analysis.

Paper	Year	Level	Dataset	Algorithm	Features	Language	Accuracy
Al-Sallab et al. [24]	2015	Sentence	ATB dataset [42]	RAE	Bag of words	MSA	-
Al-Sallab et al. [25]	2017	Sentence	ATB, 8868 tweets	RAE	Word Embedding	DA, MSA	-
Baly et al. [27]	2017	Sentence	QALB corpus (550,000 comments)	RNTN	word2vec (CBOW)	DA	-
Baly et al. [29]	2017	Sentence	ASTD	RNTN	word2vec (SG)	DA	58%
Alayba et al. [31]	2017	Sentence	Main-AHS [31] 2026 tweets	CNN	word2vec	DA, MSA	90%
Abdelhade et al. [34]	2017	Sentence	8635 stocks tweets 7440 footballs tweets	DNN	-	DA	92%
Dahou et al. [35]	2016	Sentence	ASTD, ArTwitter [43]	CNN	Word2vec	DA, MSA	85% on ArTwitter, 79% on ASTD
Al-Azani and El-Alfy [36]	2017	Sentence	ASTD	Combined LSTM	dynamic and static CBOW and SG word2vec	DA, MSA	81.63%
Alayba et al. [32]	2018	Sentence	Main-AHS, Sub-AHS, ASTD	CNN	Word Embedding	DA, MSA	92%
Heikal et al. [2]	2018	Sentence	ASTD	CNN, LSTM	Word2vec (SG)	DA, MSA	Ensemble model 65%
Elshakankery and Ahmed [37]	2019	Sentence	ArTwitter	RNN	Words weights	DA, MSA	85%
Al-Smadi et al. [38]	2018	Aspect	2291 hotels reviews	RNN	Word2vec	-	-
Al-Smadi et al. [41]	2019	Aspect	Arabic hotels reviews	LSTM	Word Embedding	DA, MSA	-
Alayba et al. [33]	2018	Sentence	Main-AHS, Sub-AHS, ASTD, ArTwitter	CNN+LSTM	Word Embedding	DA, MSA	94% Main-AHS, 95% Sub-AHS, 88% ArTwitter, 77% ASTD

3 Proposed Approach and Architecture

Figure 1 shows the overall architecture of the proposed Arabic Sentiment Analysis system using the BiLSTM deep learning model. It contains two main components: "data pre-processing and cleaning" and "Sentiment classification". In the next section, we will give all the details about each component.

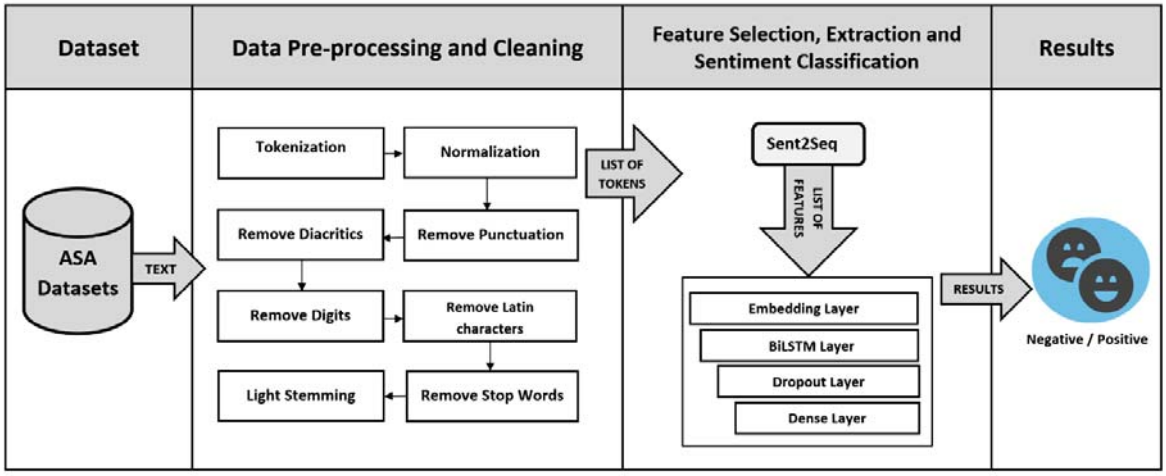

Figure 1: The architecture of proposed ASA system

3.1 Data Preprocessing and Cleaning

Sentences pre-processing, which is the first step in our method, converts the Arabic sentences to a form that is suitable for a sentiment analysis system. These pre-processing tasks include Punctuations removal, Latin characters removal, Stop word removal, Digits removal, Tokenization, Normalization, and Light Stemming. These linguistic are used to reduce the ambiguity of words in order to increase the accuracy and the effectiveness of our approach.

The pre-processing of Arabic sentences consists of the following steps:

Tokenization

Tokenization is a method for dividing texts into tokens; Words are often separated from each other by blanks (white space, semicolons, commas, quotes, and periods). These tokens could be individual words (noun, verb, pronoun, and article, conjunction, preposition, punctuation, numbers, and alphanumeric) that are converted without understanding their meaning or relationships. The list of tokens becomes an input for further processing. In this work, we use "Tokenizer" from *Keras*[1], which is the Python Deep Learning library.

1 https://keras.io/preprocessing/text/

Stop Word Removal

Stop word removal involves the elimination of insignificant words, such as (since/منذ) (for/لأجل) and (so/لذا), which appear in the sentences and do not have any meaning or indications about the content.

Other examples of these insignificant words are articles, conjunctions, pronouns (such as he/هو, she/هي, and they/هم), prepositions (such as from/من, to/الى, in/في, and about/حول), demonstratives, (such as this/هذا, these/هؤلاء, and there/اولئك), and interrogatives (such as where/أين, when/متى, and whom/لمن). Moreover, Arabic circumstantial nouns indicting time and place (such as after/بعد, above/فوق, and beside/بجانب), signal words (such as first/اولا, second/ثانيا, and third/ثالثا). A list of 202 words was prepared to be eliminated from all the sentences.

Punctuations Removal

Punctuations Removal aims to remove the punctuations symbols such as {#,-,_,.,„,;,:,‘}, because these symbols are not useful in our approach.

Latin Characters Removal and Digits Removal

In addition, we remove the Latin characters and digits, which appear in the sentences and do not have any meaning or indications in our method.

Word Normalization

Normalization aims to normalize certain letters that have different forms in the same word to one form. For example, the normalization of “ء” (Hamza),“آ” (aleph mad), “أ” (aleph with hamza on top), “ؤ”(hamza on waw), “إ” (aleph with Hamza at the bottom), and “ئ”(Hamza on ya) to “ا” (aleph). Another example is the normalization of the letter “ى” to “ي” and the letter “ة” to “ه”. We remove the diacritics such as $\{\text{ً}, \text{ٌ}, \text{ٍ}, \text{َ}, \text{ُ}, \text{ِ}, \text{ْ}\}$ because these diacritics are not used in extracting the Arabic roots and not useful in the approach proposed. Finally, we duplicate the letters that include the symbol “ّ”الشدة, because these letters are used to extract the Arabic roots, and removing them affects the meaning of the words.

Light Stemming

Light stemming is the affix removal approach that refers to a process of stripping off a small set of prefixes and/or suffixes to find the root of the word. In this work, we use the Information Science Research Institute’s (ISRI) stemmer [44]. It uses a similar algorithm to word rooting of the Khoja stemmer [45]. However, it does not employ a root dictionary for lookup. In addition, if a word cannot be rooted, it is normalized by the ISRI stemmer (*e.g.*, removing certain determinants and end patterns) instead of leaving the word unchanged. Furthermore, it defines a set of diacritical marks and affix classes. For example: (تستلزم-لزم), (مكتب-كتب)

3.2 Sentiment Classification Using Deep Learning Model

In this section, we explain the background of our sentiment classifier architecture based on the BiLSTM model. Figure 2 shows our BiLSTM model.

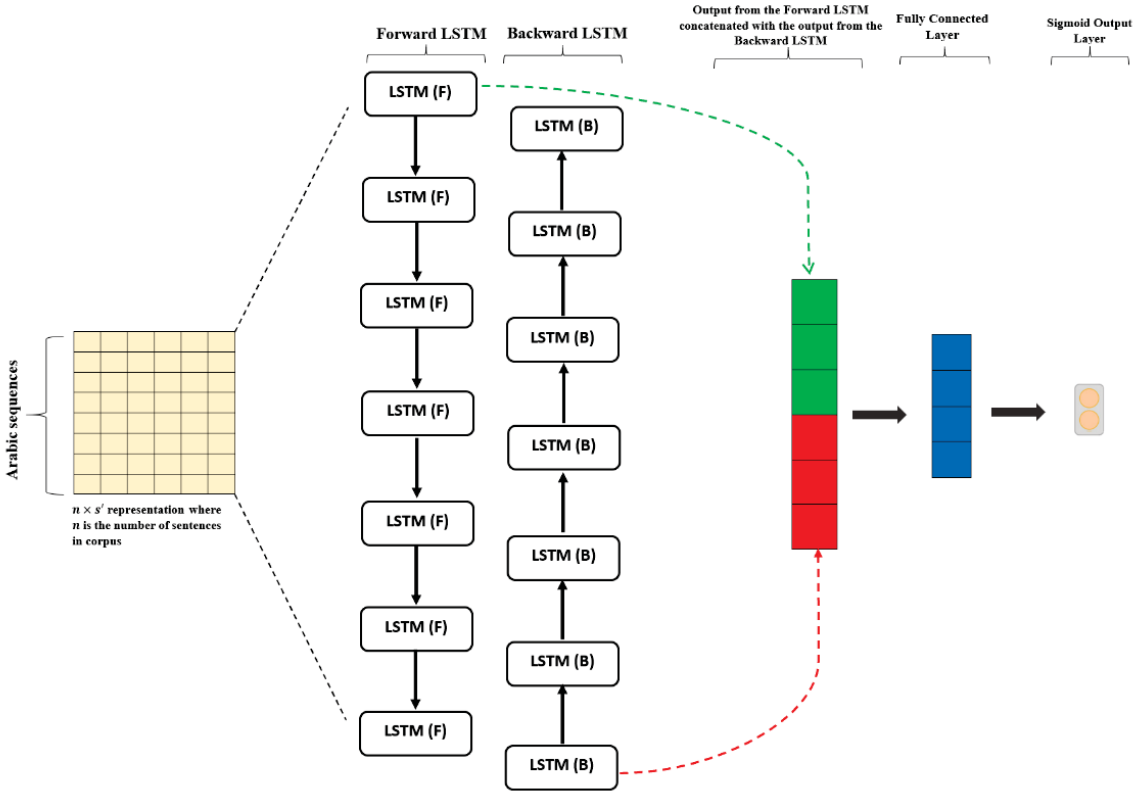

Figure 2: Our proposed BiLSTM Model

We use Keras's text pre-processing library to convert each sentence to a sequence of integers. It takes each word in the sentence and replaces it with its corresponding integer value from the vocabulary index. An entire sentence can be mapped to a vector of size s where s is the number of words in the sentence. We follow [46] zero-padding strategy such that all sentences have the same vector dimension $X \in R^{s'}$ (we chose $s' = 100$).

LSTMs are part of the recurrent neural networks (RNN) family, which are neural networks that are constructed to deal with sequential data by sharing their internal weights across the sequence [47]. LSTM addresses the problem of the vanishing error gradient and captures long term dependencies by using its gates to manage the error gradient. The Mathematical representation of LSTM can be shown as:

$$h_t = f(W_h.x_t + U_t.h_{t-1} + b_h)$$

Where x_t the current word embedding, W_h and U_t are the weights matrices, b_h is the bias term and $f(x)$ is a non-linear function, usually chosen to be $tanh$ and h_t is the regular hidden state.

$$\begin{aligned}
f_t &= \sigma(W_f.x_t + U_f.h_{t-1} + b_f) \\
i_t &= \sigma(W_i.x_t + U_i.h_{t-1} + b_i) \\
o_t &= \sigma(W_o.x_t + U_o.h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_c.x_t + U_c.h_{t-1} + b_c) \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}$$

Where i_t is called the input gate, f_t is the forget gate, c_t is the memory cell, σ is the sigmoid function, and o is the Hadamard product. Spontaneously, the forget gate decides which previous information should be forgotten, while the input gate controls what new information should be stored in the memory cell. Finally, the output gate decides the amount of information from the internal memory cell should be exposed. This gate units help a LSTM model remember significant information over multiple time steps [48]. A smaller version of the LSTM model is illustrated in Figure 3.

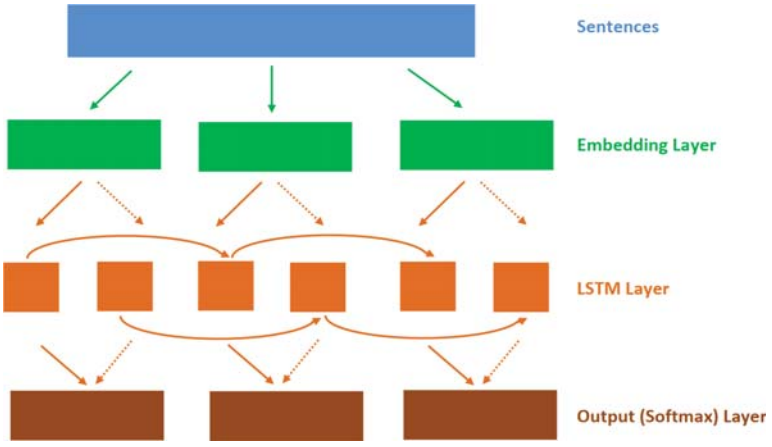

Figure 3: LSTM model architecture [48]

One drawback from the LSTM is that it does not sufficiently take into account post word information because the sentence is read only in one direction; forward. To solve this problem, we use what is known as a bidirectional LSTM, which is two LSTMs whose outputs are stacked together. One LSTM reads the sentence forward, and the other LSTM reads it backward. We concatenate the hidden states of each LSTM after they processed their respective final word [47], technically, BiLSTM applies two separate LSTM units, one for the forward direction and one for the backward direction. Two hidden states ${h_t}^{forward}$ and ${h_t}^{backward}$ from these LSTM units are concatenated into a final hidden state ${h_t}^{bilstm}$:

$$ {h_t}^{bilstm} = {h_t}^{forward} \oplus {h_t}^{backward} $$

Where $\oplus$ is concatenation operator. [49] proposed a learning model based on LSTM for semantic relationship classification and found that BiLSTM can discover richer semantic information and make full use of contextual information than LSTM. [50] utilize BiLSTM to obtain high-level semantic information features from word embedding and completes sentence-level relationship classification.

4 Experiments and Results

In this section, we will evaluate the effectiveness of our proposed method using six benchmark sentiment analysis datasets. Section 4.1 presents the data sets. Section 4.2 describes the evaluation metrics. Section 4.3 details the parameters setting. Finally, Section 4.4 presents the experimental results.

4.1 Datasets

In this paper, the experiments are conducted using six benchmarks sentiments analysis datasets. Furthermore, to show the flexibility of our model on various domains. We have used only two sentiment classes, *i.e.*, Positive and Negative, and we removed the objective class because the class distribution was highly skewed, and it is more important to focus on opinion classification rather than subjectivity classification.

4.1.1 ASTD: Arabic Sentiment Tweets Dataset

The authors of [30] presented a sentiment analysis dataset from Twitter, they were grouped into four categories (positive, negative, neutral, and objective).

4.1.2 ArTwitter: Twitter Data set for Arabic Sentiment Analysis

The authors of [43] have manually built a labeled sentiment analysis dataset form Twitter. The dataset contains 2000 labeled tweets (1000 positive tweets and 1000 negative ones) collected by using a tweet crawler.

4.1.3 LABR: A Large-Scale Arabic Book Reviews Dataset

In [51], they presented a large dataset of Arabic Book reviews. This dataset contains over 63,000 book reviews in Arabic. The book reviews were harvested from the website[2] during March 2013. Each book review comes with a review ID, the user ID, the book ID, the rating (1 to 5), and the text of the review.

4.1.4 MPQA: Multi-Perspective Question Answering

The authors of [52] presented a news articles dataset. The dataset contains news articles from a wide variety of news sources manually annotated for opinions and other private states (*i.e.*, beliefs, emotions, sentiments, speculations, etc.)

4.1.5 Large Arabic Multi-domain Resources for Sentiment Analysis

[53] proposed a dataset of 33K automatically annotated Reviews in Domains of Movies, Hotels, Restaurants, and Products. The datasets cover four domains as follows:

i. **Hotel Reviews (HTL):** For the hotel's reviews were scrapped from TripAdvisor[3].
ii. **Restaurant Reviews (RES):** For the restaurant's reviews were scarped from Qaym[4] and TripAdvisor.
iii. **Movie Reviews (MOV**): The movie's domain dataset was built out of scrapping reviews from Elcinemas[5] covering around 1K movies.
iv. **Product Reviews (PROD):** For the Products domain, a dataset of reviews was scraped from the Souq[6] website. The dataset includes reviews from Egypt, Saudi Arabia, and the United Arab Emirates.

4.1.6 Arabic Health Services Dataset

The authors of [31] presented an Arabic analysis dataset collected from Twitter. It has two classes (positive and negative). The dataset contains 2026 tweets, and it is an unbalanced dataset that has 1398 negative tweets and 628 positive tweets. We keep it "Main-AHS" as their authors call it.

Table 2 shows the statistics of each dataset used in the ASA system.

2 www.goodreads.com
3 www.tripadvisor.com
4 www.Qaym.com
5 http://www.Elcinemas.com
6 http://www.souq.com

Table 2: The dataset used to sentiment classification

Dataset	Positive	Negative	***Total***
ASTD	777	812	**1589**
ArTwitter	993	958	**1951**
LABR	8224	8224	**16448**
MPQA	4597	5399	**9996**
Multi-Domain Resource	24948	6650	**31598**
Main-AHS	628	1398	**2026**
Total	**39539**	**22043**	**61582**

4.2 Evaluation metrics

To evaluate the performance of the sentiment analysis system, we have employed four well-known metrics, namely, Accuracy, Precision, Recall, and F-measure. There are widely used to measure sentiment prediction.

Accuracy, Precision, Recall, and F-measure are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$F1 - measure = \frac{2.precision}{precision + recall}$$

Where TP is the number of sentences that are positive and predicted correctly as positive, FP is the number of sentences that are negative and predicted incorrectly as positive, TN is the number of sentences that are negative and predicted correctly as negative, and FN is the number of sentences that are positive and predicted incorrectly as negative. Note that, the higher the Precision, the more accurate the prediction of the positive class, A high recall means a high number of sentences from the same class is labeled to its exact class, F1-measure is a weighted average of Precision and recall, and for the accuracy, it simply reports the ratio of the correctly classified sentences regardless of their class.

Besides, A comprehensive evaluation of classifier performance can be obtained by the ROC:

$$ROC = \frac{P(x/Positive)}{P\left(x/Negative\right)}$$

Where $P(x/C)$ denotes the conditional probability that a data entry has the class label C. A ROC curve plots the classification results from the most positive to the most negative classification [54].

4.3 Parameters setting

There are a number of parameters to tune. In this work, several experiments were conducted to find the optimal parameters. Only the parameters that yield the best results are reported. For this purpose, we use the size of the training set is 80% of the whole dataset, and the test set contains the remaining 20% of the dataset. The number of epochs is 10 for all the experiments. For the regularization of the neural networks and to avoid the over-fitting problem, we apply Dropout, with a dropout rate of 0.5.

4.4 Experimental Results

This section is intended to compare the performance of the proposed model described in Section 3 with the state-of-art Arabic Sentiment Analysis methods. Additionally, we compare our results of a deep learn-

ing model with two baselines prevalent in traditional machine learning, namely the Random Forest classifier and the Support Vector Machine classifier. Also, we have mentioned the impact of Light stemming on our proposed approach.

Table 3 presents the comparison between our proposed approach and the state-of-art Arabic Sentiment Analysis methods on each dataset. It is clear that our deep learning model improved the performance of sentiment classification, and our model achieves an Accuracy of 72.25% on the ASTD dataset, 91.82% on the ArTwitter dataset, and 92.61% on Main-AHS dataset, which outperforms the state-of-the-art methods.

Table 4 and Figure 4 shows the detailed performance of our deep learning model on each of the six Arabic datasets. We compare our results of the BiLSTM model with two baselines prevalent in traditional machine learning, namely the Random Forest classifier and the Support Vector Machine classifier. Our results in Table 4 consistently reveal superior performance through the use of our BiLSTM model over the baseline traditional machine learning.

Table 3: Performance comparison with state-of-art methods on the same dataset

dataset	Approach	Technique	Accuracy (%)	F1-measure (%)
ASTD	Heikal *et al.* [2]	CNN (fully connected layer size = 100)	64.30	64.09
		LSTM (dropout rate = 0.2)	64.75	62.08
		Ensemble model	65.05	64.46
	Baly *et al.* [29]	RNTN	58.50	53.6
	Alayba *et al.* [33]	CNN+LSTM	77.62	-
	Our work	**BiLSTM**	**79.25**	**76.83**
ArTwitter	Elshakankery and Ahmed [37]	RNN	85.00	
	Dahou *et al.* [35]	CNN	85.01	
	Alayba *et al.* [33]	CNN+LSTM	88.10	-
	Al-Azani and El-Alfy [36]	Combined LSTM	87.27	-
	Our work	**BiLSTM**	**91.82**	**92.39**
Main-AHS	Alayba *et al.* [32]	CNN	92.00	-
	Alayba *et al.* [31]	CNN	90.00	-
	Our work	**BiLSTM**	**92.61**	**86.03**

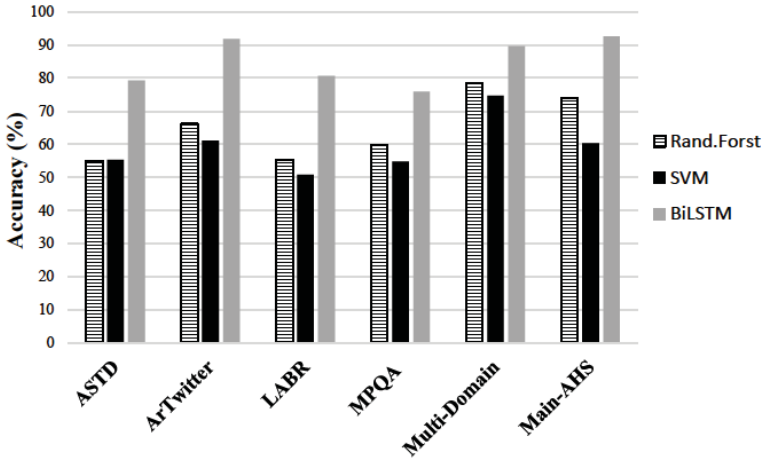

Figure 4: Results of Accuracy for the proposed approach on each dataset

Table 4: Results of precision, recall, and F1-measure for the proposed system on each dataset

		Baseline: Traditional Machine Learning		Our Deep Learning model
Dataset	Metrics	Rand.Forst	SVM	BiLSTM
ASTD	Precision	55.58	55.32	**83.10**
	Recall	55.13	55.30	**72.35**
	F1-measure	54.10	55.27	**76.83**
ArTwitter	Precision	66.34	67.34	**93.30**
	Recall	66.43	58.31	**91.64**
	F1-measure	66.21	53.65	**92.39**
LABR	Precision	55.27	50.88	**79.79**
	Recall	55.23	50.88	**80.04**
	F1-measure	55.20	50.72	**79.41**
MPQA	Precision	58.83	54.80	**74.74**
	Recall	58.33	54.86	**68.41**
	F1-measure	58.23	54.67	**70.72**
Multi-Domain	Precision	59.89	51.74	**91.89**
	Recall	50.84	50.78	**95.43**
	F1-measure	46.54	49.33	**93.52**
Main-AHS	Precision	83.34	56.02	**95.90**
	Recall	62.31	56.08	**80.28**
	F1-measure	61.76	56.04	**86.03**

Our motivation behind choosing the BiLSTM-based approach such a classification based-method, due to the Forward-Backward encapsulate contextual information during varied Arabic based-target learning stages. More generally, we have conducted other experiments by applying two other deep learning methods with regard to specific training settings. We compare our results of BiLSTM with two popular deep learning models, such as the Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) network. [46] defined CNNs to have convolving filters over each input layer in order to generate the best features. [55] confirmed that CNN is a powerful tool to select features in order to improve the prediction accuracy. [56] showed the capabilities of LSTMs in learning data series by considering the previous outputs.

Table 5: Comparison of accuracy results of different deep learning models

	Dataset	**Deep Learning model**		
		BiLSTM	LSTM	CNN
Without Stemming	**ASTD**	**75.89**	75.26	74.21
	ArTwitter	**87,21**	86.19	85,68
	LABR	79.33	79,27	**79,64**
	MPQA	**73,90**	71,25	69,40
	Multi-Domain	88,64	88,24	**88,92**
	Main-AHS	**90.64**	87.93	90.39
Arabic Light Stemming (ISRI Stemmer)	**ASTD**	**79.25**	77.57	79.04
	ArTwitter	**91.82**	90.54	90.79
	LABR	80.70	78.51	**81.61**
	MPQA	**75.85**	73.60	70.70
	Multi-Domain	**89.70**	89.57	87.69
	Main-AHS	**92.61**	89.90	90.39

As demonstrated in Table 5 We show that the better performance obtained in terms of Accuracy is obtained by our BiLSTM model on the majority of the benchmark datasets used in the learning stage.

Furthermore, The results in Table 5 have shown that applying sentiment classification with light stemming gives a significant performance improvement of 3.36% on ASTD, 4.61% on ArTwitter, 1.37% on LABR, 1.95% on MPQA, 1.06% on Multi-Domain, and 1.97% on Main-AHS dataset.

Additionally, In Figure 5 we calculate the ROC score for each baseline traditional machine learning method and Deep Learning model on each dataset used in the experiment. As shown, ROC scores of baseline traditional machine learning (SVM and Random forest) generally have the lowest ROC score in the different datasets compared with the other deep learning models in general and BiLSTM based model in particular.

Furthermore, Figure 6, Figure 7, and Figure 8. illustrate the accuracies on different datasets over 10 epochs. Each line represents a different deep learning model. The Accuracy results confirmed our finding

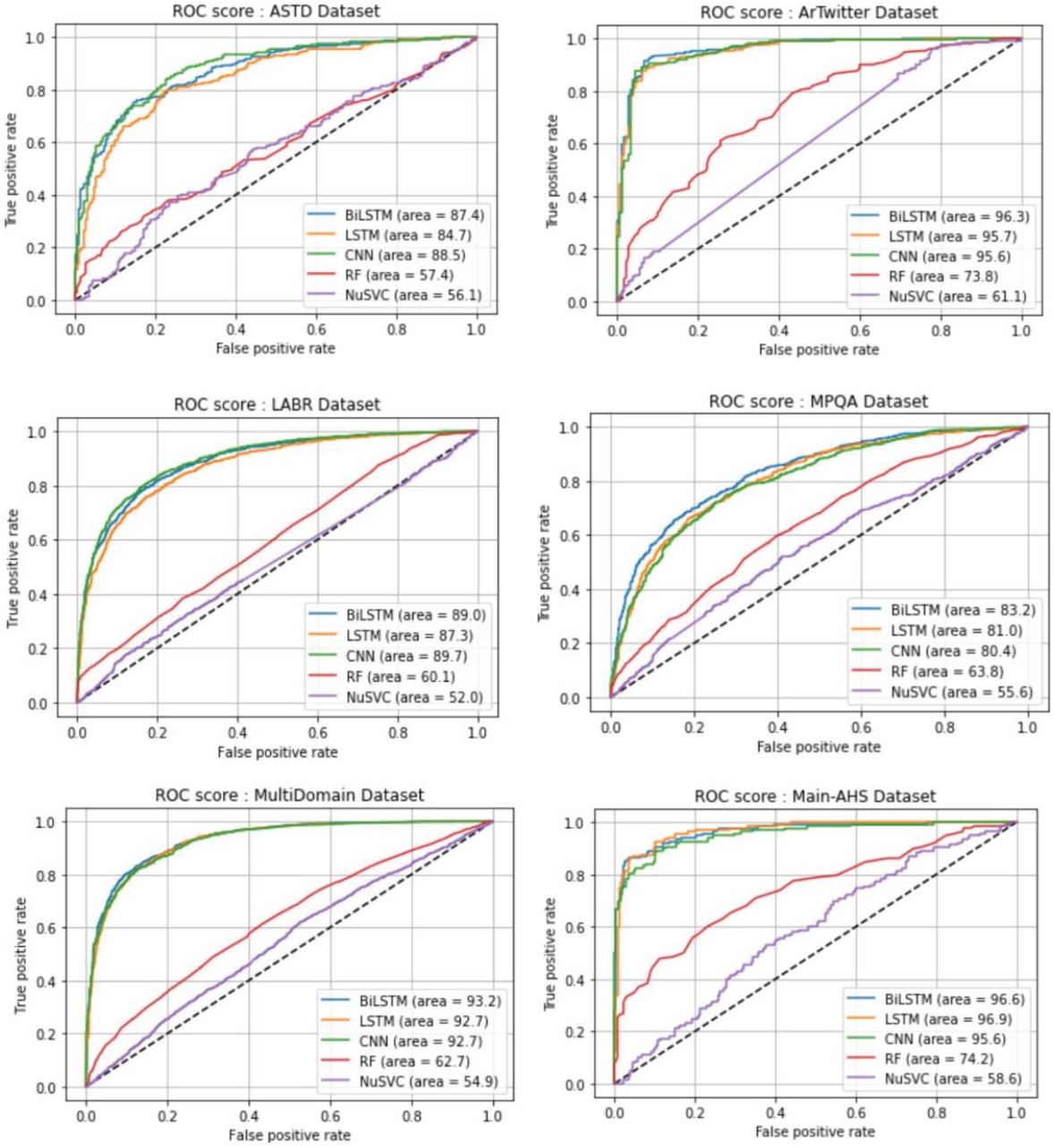

Figure 5: ROC score on each dataset

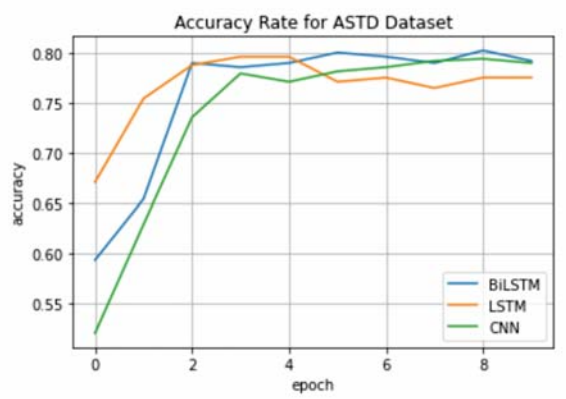

Figure 6: Accuracy on the test set for ASTD dataset using different deep learning model

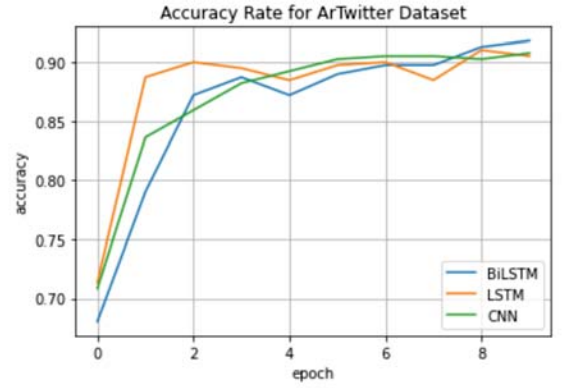

Figure 7: Accuracy on the test set for ArTwitter dataset using different deep learning model

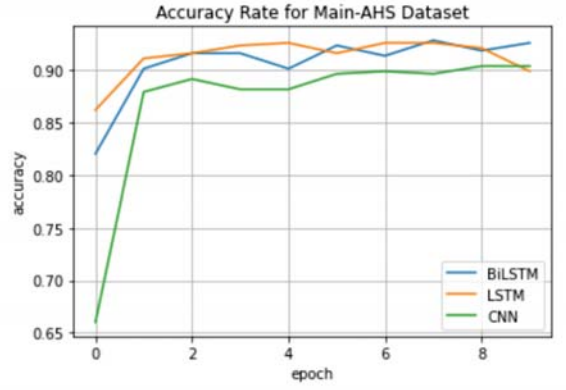

Figure 8: Accuracy on the test set for Main-AHS dataset using different deep learning model

and quoted that the BiLSTM achieves better results than other deep learning models on the majority of the benchmark datasets.

5 Discussion

The objective of this work is to propose a novel Arabic Sentiment analysis approach to overcome the limited ability of the feed-forward model by extracting unlimited contextual information on both directions of the Arabic sentence. From the results presented in Section 4.4, we can highlight that our proposed method is able to yield the best results in terms of sentiment prediction quality. BiLSTM shows higher results than other deep learning models (CNN and LSTM) on the majority of the benchmark datasets. This is due to the fact that BiLSTM can more effectively learn the context of each word in the text, it accesses both preceding and succeeding contextual features by combining a forward hidden layer and a backward hidden layer. Moreover, we found that BiLSTM can discover richer semantic information and make full use of contextual information than LSTM.

In Addition, Our results through the use of our BiLSTM model over the baseline traditional machine learning. This due to the fact that Deep Learning algorithms achieved accuracies better than traditional machine learning methods, and according to [57], large training data makes SVM inefficient and costly, as SVM is not scalable to huge size data. When the training data is noisy and imbalanced, it can affect the outcome of SVM due to its high training execution and low generalization error [57]. For the Random Forest algorithm, the complexity grows with the number of trees in the forest and the number of training samples we have.

Furthermore, our model was enriched with morphological features, including stems, in order to overcome the lexical sparsity and ambiguity issue. The results show that applying light stemming gives a significant improvement of 2.39% as an average on the six datasets used in our experiments, and according to [6], these features achieved significant performance improvements on the data containing a mixture of MSA and DA.

The main advantages of this work are the capacity to effectively improve the quality of sentiment predictions by investigating the benefits of Arabic pre-processing such as tokenization, Punctuations removal, Latin characters removal, Digits removal, Normalization, Light Stemming, and the ability to consider the contextual information by dealing with both forward and backward dependencies.

6 Conclusion and Future work

In this work, we have addressed the Sentiment Analysis problem for Arabic Text. This paper exploits the benefit of using a deep learning model on the performance of the Arabic Sentiment Analysis system. We have used the BiLSTM deep learning model with the ability of extracting the contextual information to predict the sentiment of the Arabic text. Experiments are conducted on six benchmark datasets to evaluate the performance of our presented approach. The results show the effectiveness of BiLSTM to perform sequential data models and can further extract the contextual information by dealing with both forward and backward dependencies from the feature sequences. Comparisons with some state-of-art baseline methods, it demonstrates that in most cases our deep learning model is more effective and efficient in terms of the classification quality. Besides, the model achieves significant improvements in the Accuracy and F1-measure results over the existing models. Our Model will definitely help to ensure further exploration.

In future work, we plan to compare the impact of different recent contextualized word embedding (*e.g.*, GoVe, ELMO, ULMFiT, BERT, GPT, GPT-2, and XLNet) on the performance of our presented deep learning model using different Arabic sentiment analysis datasets. Furthermore, we plan to work on dialectal Arabic corpora to cover all variations of Arabic words.

References

[1] M. HAMMAD, Mustafa et AL-AWADI, "Sentiment analysis for arabic reviews in social networks using machine learning," in *In Information Technology: New Generations*, 2016, pp. 131–139.

[2] M. Heikal, M. Torki, and N. El-Makky, "Sentiment Analysis of Arabic Tweets using Deep Learning," *Procedia Comput. Sci.*, vol. 142, pp. 114–122, 2018, doi: 10.1016/j.procs.2018.10.466.
[3] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," *2013 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2013*, pp. 271–276, 2013, doi: 10.1109/ICICES.2013.6508366.
[4] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis," *2015 IEEE 2nd Int. Conf. Recent Trends Inf. Syst. ReTIS 2015 - Proc.*, vol. 4, no. 2, pp. 353–358, 2015, doi: 10.1109/ReTIS.2015.7232904.
[5] M. Korayem, D. Crandall, and M. Abdul-Mageed, "Subjectivity and Sentiment Analysis of Arabic: A Survey," in *In International conference on advanced machine learning technologies and applications*, 2012, vol. 322, no. December, pp. 128–139, doi: 10.1007/978-3-642-35326-0.
[6] G. Badaro *et al.*, "A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 3, 2019, doi: 10.1145/3295662.
[7] S. HOCHREITER and J. SCHMIDHUBER, "Long Short-Term Memory," *Neural Comput.*, vol. 1780, pp. 1735–1780, 1997.
[8] A. Graves, A. Mohamed, and G. Hinton, "SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton Department of Computer Science, University of Toronto," no. 3, pp. 6645–6649, 2013.
[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 3156–3164, 2015, doi: 10.1109/CVPR.2015.7298935.
[10] D. Eck and J. Schmidhuber, "A First Look at Music Composition using LSTM Recurrent Neural Networks," *Idsia*, pp. 1–11, 2002, [Online]. Available: http://people.idsia.ch/~{}juergen/blues/IDSIA-07-02.pdf%0Ahttp://www.idsia.ch/~{}juergen/blues/IDSIA-07-02.pdf.
[11] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Inf. Process. Manag.*, vol. 56, no. 2, pp. 320–342, 2019, doi: 10.1016/j.ipm.2018.07.006.
[12] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *Proceedings of the 2012 International Conference on Collaboration Technologies and Systems, CTS 2012*, 2012, pp. 546–550, doi: 10.1109/CTS.2012.6261103.
[13] M. Al-Kabi, A. Gigieh, I. Alsmadi, H. Wahsheh, and M. Haidar, "The fourth international conference on information and communication systems," in *The fourth international conference on information and communication systems*, 2013, pp. 23–25.
[14] A.-K. M. Abdulla NA, Al-Ayyoub M, "No Title," in *Int J Big Data Intell*, 2014, p. (1–2):103.
[15] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 20–37, 2014, doi: 10.1016/j.csl.2013.03.001.
[16] R. Duwairi and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text," *J. Inf. Sci.*, vol. 40, no. 4, pp. 501–513, 2014, doi: 10.1177/0165551514534143.
[17] A.-S. A. Al-Subaihin AA, Al-Khalifa HS, "19.," in *Proceedings of the 13th international conference on information integration and web-based applications and services.*, 2011, pp. 543–546.
[18] Ayyoub, Essa, and Alsmadi, "Lexicon-based sentiment analysis of Arabic tweets Mahmoud Al-Ayyoub * and Safa Bani Essa Izzat Alsmadi," *Int. J. Soc. Netw. Min. 2(2) 101*, vol. X, 2015, doi: 10.1504/IJSNM.2015.072280.
[19] A. El-halees, "Arabic opinion mining using combined classification approach," *Proceeding Int. Arab Conf. Inf. Technol. Azrqa, Jordan.*, pp. 264–271, 2011.
[20] T. H. Soliman, M. A. Elmasry, A. Hedar, and M. M. Doss, "Sentiment Analysis of Arabic Slang Comments on Facebook," *Int. J. Comput. Technol.*, vol. 12, no. 5, pp. 3470–3478, 2014, doi: 10.24297/ijct.v12i5.2917.
[21] I. S. El-Makky N, Nagi K, El-Ebshihy A, Apady E, Hafez O, Mostafa S, "No Title," in *The 3rd ASE international conference on social informatics (SocialInformatics 2014)*, 2015.
[22] R. Duwairi, "No Title," in *In: Information and communication systems (ICICS)*, 2015, pp. 166–170.
[23] L. Deng and Y. Dong, "Foundations and trends® in signal processing," *Signal Processing*, vol. 7, pp. 3–4, 2014.
[24] A. Al Sallab, H. Hajj, G. Badaro, R. Baly, W. El-Hajj, and K. Shaban, "Deep learning models for sentiment analysis in Arabic," in *Proceedings of the second workshop on Arabic natural language processing*, 2015, pp. 9–17.
[25] A. Al-Sallab, R. Baly, H. Hajj, K. B. Shaban, W. El-Hajj, and G. Badaro, "AROMA: A recursive deep learning model for opinion mining in Arabic as a low resource language," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 4, 2017, doi: 10.1145/3086575.
[26] R. Baly *et al.*, "OMAM at SemEval-2017 Task 4: Evaluation of English State-of-the-Art Sentiment Analysis Models for Arabic and a New Topic-based Model," pp. 603–610, 2018, doi: 10.18653/v1/s17-2099.
[27] R. Baly, H. Hajj, N. Habash, K. B. Shaban, and W. El-Hajj, "A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 4, 2017, doi: 10.1145/3086576.
[28] B. Mohit, A. Rozovskaya, N. Habash, W. Zaghouani, and O. Obeid, "The first QALB shared task on automatic text correction for Arabic," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 39–47.
[29] R. Baly *et al.*, "A Characterization Study of Arabic Twitter Data with a Benchmarking for State-of-the-Art Opinion Mining Models," pp. 110–118, 2017, doi: 10.18653/v1/w17-1314.

[30] M. Nabil, M. Aly, and A. Atiya, "Astd: Arabic sentiment tweets dataset," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2515–2519.
[31] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," pp. 114–118, 2017, doi: 10.1109/asar.2017.8067771.
[32] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Improving sentiment analysis in Arabic using word representation," in *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, 2018, pp. 13–18.
[33] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "A combined CNN and LSTM model for arabic sentiment analysis," in *International cross-domain conference for machine learning and knowledge extraction*, 2018, pp. 179–191.
[34] N. Abdelhade, T. H. A. Soliman, and H. M. Ibrahim, "Detecting Twitter users' opinions of Arabic comments during various time episodes via deep neural network," in *International Conference on Advanced Intelligent Systems and Informatics*, 2017, pp. 232–246.
[35] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, "Word embeddings and convolutional neural network for arabic sentiment classification," in *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, 2016, pp. 2418–2427.
[36] S. Al-Azani and E.-S. M. El-Alfy, "Hybrid deep learning for sentiment polarity determination of arabic microblogs," in *International Conference on Neural Information Processing*, 2017, pp. 491–500.
[37] K. Elshakankery and M. F. Ahmed, "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis," *Egypt. Informatics J.*, vol. 20, no. 3, pp. 163–171, 2019.
[38] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," *J. Comput. Sci.*, vol. 27, pp. 386–393, 2018.
[39] M. Althobaiti, U. Kruschwitz, and M. Poesio, "AraNLP: A Java-based library for the processing of Arabic text," 2014.
[40] A. Pasha *et al.*, "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic.," in *LREC*, 2014, vol. 14, pp. 1094–1101.
[41] M. Al-Smadi, B. Talafha, M. Al-Ayyoub, and Y. Jararweh, "Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 8, pp. 2163–2175, 2019.
[42] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The penn arabic treebank: Building a large-scale annotated arabic corpus," in *NEMLAR conference on Arabic language resources and tools*, 2004, vol. 27, pp. 466–467.
[43] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, 2013, pp. 1–6.
[44] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, 2005, vol. 1, pp. 152–157.
[45] S. Khoja and R. Garside, "Stemming arabic text," *Lancaster, UK, Comput. Dep. Lancaster Univ.*, 1999.
[46] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv Prepr. arXiv1408.5882*, 2014.
[47] M. Cliche, "Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms," *arXiv Prepr. arXiv1704.06125*, 2017.
[48] H. Ghulam, F. Zeng, W. Li, and Y. Xiao, "Deep learning-based sentiment analysis for roman urdu text," *Procedia Comput. Sci.*, vol. 147, pp. 131–135, 2019.
[49] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," *arXiv Prepr. arXiv1512.01100*, 2015.
[50] P. Zhou *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2016, pp. 207–212.
[51] M. Aly and A. Atiya, "Labr: A large scale arabic book reviews dataset," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, pp. 494–498.
[52] V. Stoyanov, C. Cardie, and J. Wiebe, "Multi-perspective question answering using the OpQA corpus," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 2005, pp. 923–930.
[53] H. ElSahar and S. R. El-Beltagy, "Building large arabic multi-domain resources for sentiment analysis," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2015, pp. 23–34.
[54] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, 2006, pp. 1015–1021.
[55] B. Athiwaratkun and K. Kang, "Feature representation in convolutional neural networks," *arXiv Prepr. arXiv1507.02313*, 2015.
[56] F. A. Gers, D. Eck, and J. Schmidhuber, "Applying LSTM to time series predictable through time-window approaches," in *Neural Nets WIRN Vietri-01*, Springer, 2002, pp. 193–200.
[57] P. Koturwar, S. Girase, and D. Mukhopadhyay, "A survey of classification techniques in the area of big data," *arXiv Prepr. arXiv1503.07477*, 2015.