ISSN: 0067-2904

A Review for Arabic Sentiment Analysis Using Deep Learning

Anwar Abdul-Razzaq Hussien*, Nada A. Z. Abdullah
Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq.

Received: 27/8/2022 Accepted: 27/1/2023 Published: 30/12/2023

Abstract

Sentiment Analysis is a research field that studies human opinion, sentiment, evaluation, and emotions towards entities such as products, services, organizations, events, topics, and their attributes. It is also a task of natural language processing. However, sentiment analysis research has mainly been carried out for the English language. Although the Arabic language is one of the most used languages on the Internet, only a few studies have focused on Arabic language sentiment analysis.

In this paper, a review of the most important research works in the field of Arabic text sentiment analysis using deep learning algorithms is presented. This review illustrates the main steps used in these studies, which include pre-processing, feature extraction and classification, as well as the datasets used. In the end, all the research works are compared in terms of their methodology and results. The findings demonstrated that the majority of deep learning models, including CNN and LSTM, outperformed many of the machine learning models analyzed, and that the size of the training datasets had a direct correlation with the model's performance. Where larger datasets resulted in more successful model training.

Keywords: Arabic Sentiment Analysis, Deep Learning, Neural Networks, Modern Standard Arabic, Dialect Arabic.

مراجعة لتحليل المشاعر في النصوص العربية باستخدام التعلم العميق

أنوار عبدالزراق حسين*, ندا عبدالزهرة عبدالله
قسم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

تحليل المشاعر هو مجال بحثي يحلل الرأي البشري والمشاعر والتقييم والعواطف تجاه الكيانات مثل المنتجات والخدمات والمنظمات والأحداث والموضوعات وخصائصها وهي مهمة معالجة اللغات الطبيعية. ومع ذلك، فقد تم إجراء أبحاث تحليل المشاعر بشكل أساسي للغة الإنجليزية. على الرغم من أن اللغة العربية هي واحدة من أكثر اللغات استخدامًا على الإنترنت، إلا أن دراسات قليلة فقط ركزت على تحليل المشاعر باللغة العربية.

يستعرض هذا البحث مراجعة لأهم الأعمال البحثية في مجال تحليل آراء النص العربي باستخدام خوارزميات التعلم العميق، كما توضح هذه المراجعة الخطوات الرئيسية المستخدمة في هذه الأبحاث والتي تشمل: المعالجة المسبقة، واستخراج السمات والتصنيف، بالإضافة إلى مجموعة البيانات المستخدمة. في النهاية، تتم مقارنة جميع الأعمال البحثية من حيث المنهجية والنتائج. أظهرت النتائج أن غالبية نماذج التعلم العميق، بما في ذلك CNN

*Email: Anwar.Abdulrazzaq1201a@sc.uobaghdad.edu.iq

وLSTM، تفوقت على العديد من نماذج التعلم الآلي التي تم تحليلها، وأن حجم مجموعات بيانات التدريب كان له علاقة مباشرة بأداء النموذج. حيث أدت مجموعات البيانات الأكبر إلى تدريب نموذج أكثر نجاحًا.

1. Introduction

With the rise in Internet use, social networks have spread allowing users to connect everywhere and anytime, and therefore their daily lives became depended on it [1].
People use the internet to express their sentiments about what is happening in society and the market in terms of campaigns and products as well as political movements. Thus, both the business world and some organizations as well as the scientific community, are becoming more interested in extracting these sentiments. As a result, several fields of research have emerged, including Sentiment Analysis (SA), which is based on information mining and human-computer interaction [2].

SA is considered a research area in progress. It is one of the natural language processing applications, also known as opinion analysis, opinion mining, sentiment mining, subjectivity analysis, impact analysis, emotion analysis, revision extraction, etc. Sentiment Analysis explores the writer's sentiments, categorizing texts based on the feelings they contain and analyzing their contextual polarity into positive, negative, or neutral [3].

Because of the importance of this topic, and for many years, many researchers conducted studies that used machine learning to analyze sentiments in texts in various languages such as English, Chinese, and Arabic, but recently deep learning has been used. The Arabic language was less researched than the rest of the languages in terms of the number of studies, even though more than 330 million people uses it. This is due to the complexity of its morphological structure and its diversity, as it is in standard or colloquial. It also contains many dialects that differ from one country to another and sometimes within the same country, as well as the lack of tools for deep learning of the Arabic language [4].

The remainder of the paper will have the following structure: the second section will explore a review of the research on Arabic sentiment analysis using deep learning. The third section will illustrate the characteristics of the Arabic language. The fourth section will explain the sentiment analysis model, which includes subsections, and the fifth section will present a brief comparison of the annotated research. The conclusion and future work will be in the last section.

2. Literature Survey

Many researchers have conducted studies on sentiment analysis in texts written in many languages using many approaches, the most important of which is machine learning. Many mechanisms of ML were used in these studies, such as classical lexical approaches, shallow machine learning and more recently deep learning.

This section summarizes the most important studies that have were conducted to analyze sentiment in Arabic texts focusing on those using deep learning approaches.
The first study on Arabic Sentiment Analysis published was by Al Sallab et al in 2015 [5], where they introduced four models. The first three models were Deep Neural Network (DNN), Combined Deep Belief Network (DBN), and Deep Auto Encoder (DAE). The input feature vectors based on the Bag of Words with features based on ArSenl [6] were employed as well as other standers lexicon features, with the observation that the vector length is fixed. The fourth model was based on the Recursive Auto Encoder (RAE), and it aimed to present a solution to the first three models' lack of context management, and the length of the feature vector was variable because the vector was the raw word indices that comprise each sentence. The RAE model outperformed other models in the experiment by reaching 74% accuracy.

Dahou et al. in 2016 [7] employed a web-crawler to collect big croups for Arabic text in both Modern Standard Arabic (MSA) and dialectal consisting of 10 billion words. They chose 3.4 billion words of them to examine two word embedding architectures Continuous Bag Of Words (CBOW) and Skip-Grams (SG). Then, by using a pre-trained word representation, a Convolutional Neural Network (CNN)-based model was trained to classify sentiments.

Al Sallab et al. in 2017 [8] introduced improvements to the RAE model that they previously proposed in 2015. To improve representation and address the morphological complexity of the Arabic text and overfitting. The authors suggested morphological tokenization and merging sentiment and semantic embeddings.

Three datasets were used to evaluate the model. The test results showed that it outperformed the old model with accuracy reaching 86%. They also contributed with their improved RAE model in Sem-Eval 2017 task 4 [9] and obtained 41% accuracy of the prediction.

Abdelhade et al. in 2018 [10] employed the DNN model, which consists of eight hidden layers with a final softmax layer used to identify the class label of the input text, to predict the polarity of Arabic datasets collected from Twitter in two different domains. Tests showed the model successfully exceeded the baseline with an accuracy of 94%.

Baly et al. in 2017 [11] contributed first to the creation of the first morphologically and orthographically enriched Arabic language, which is the Arabic Sentiment Treebank (ARSENTB), and then used it to train the Recursive Neural Tensor Network (RNTN) model. They embedded the input into vectors by using word2vec embedding using the CBOW model. The result of the experiments showed that the accuracy of the RNTN model reached 80%.

Another study [12] conducted by the same authors where they trained the RNTN model on a different dataset consisting of 10,006 tweets. The model was trained twice, first using lemmas, and the other using raw data. The accuracy of the first one outperformed on the other.

Alayba et al. [13] collected a dataset from Twitter on the health services topic. They used the word2vec model to embed the input layer, and for classifying the sentiment of the tweet, DNN and CNN models were used. Despite having been trained on a relatively small dataset, the CNN model had the highest accuracy.

Al-Azani et al. [14] employed two deep models, CNN, and Long Short-Term Memory (LSTM), to investigate various models to classify the sentiment of Arabic text in microblogs: CNN, CNN-LSTM, simple LSTM, stacked LSTM, and combined LSTM. For word embedding, they utilized two embedding techniques (CBOW and SG) with static and dynamic initializations. The model test results showed that the combined LSTM model trained using dynamic CBOW achieved higher accuracy than the other models.

The authors of [13] suggested another model in [15] for Arabic sentiment analysis for the same dataset that they previously collected and used. They employed different Arabic corpus to construct Word2Vec models that were used to train a CNN model and lexicon model to overcome the limitation of training on a small dataset. The accuracy of the proposed model outperformed the previous one, and reached 92% and 95% for Main-AHS and Sub-AHS, respectively.

Also, the same authors investigated the merit of combining the CNN and LSTM models to improve the sentiment analysis of the Arabic datasets [16]. Because of the difficulties of morphology and orthography in Arabic, the effectiveness of applying multiple levels of word embeddings were also investigated: character level, word-level, and Ch5gram-level. Where the input is sent first to a CNN layer and subsequently to an LSTM layer. The output layer then uses a sigmoid function to make the final prediction. Due to the experiments, an improvement was observed in the accuracy of sentiment classification for the Main-AHS dataset and Sub-AHS dataset, where it reached 0.9424 and 0.9568, respectively.

Heikal et al. [17] built an ensemble model that integrated a CNN and a bi-directional LSTM (BILSTM) to predict the sentiment of the text in Arabic tweets on the Arabic Sentiment Tweets Dataset (ASTD). They conducted independent experiments using the ensemble model, CNN, and BILSTM. The result of the experiments showed that the ensemble model had high accuracy and F1-score compared with the rest, which reached 64.46% and 65.05%, respectively.

Wint et al. [18] built a hybrid deep model combining two CNN and BILSTM models and called it (H2CBi). They used two different kinds of social network services (SNS), product review data which are Movie Review, Amazon and Yelp, and social network site data, which are, Yelp, Twitter I, Twitter II, Facebook and FormSpring.me. They employed the CNN model to extract the features from the data whether they were positive or negative, bullied or not bullied, and the BILSTM to create a representation at the sentence level. For the word representation, they used word2vec, fastText and Glove pre-trained word embedding. Due to the experiments conducted on the seven datasets, the results showed that three of six models of H2CBi outperformed the baseline result on six of the datasets.

Mohammed A. and Kora R. [19] presented a corpus of forty thousand texts written in MSA and Egyptian dialects. They proposed three models to analyze the polarity of the tweet in this corpus, which were CNN, LSTM, and recurrent convolution neural network (RCNN). They conducted experiments to evaluate the three models, and according to the results, LSTM performed better than the other two models with accuracy reaching 81.31%. Also, the accuracy reached 88.05% by applying the augmentation technique to the corpus.

Elzayady et al. [20] employed three deep learning models for sentiment analysis on two datasets, Hotels Reviews (HTL) and Book Reviews (LABR), which were CNN, LSTM, and a combined CNN model with the Recurrent Neural networks (RNNs) model (CNN-LSTM). The experimental results showed that the combined model had the highest average accuracy with 85,83% for HTL and 86,88% for LABR compared with CNN and LSTM separated.
Al-Bayati et al. [21] investigated the LSTM model to enhance the accuracy of classifying the sentiment of the Arabic text into positive and negative. They used a word embedding layer as a hidden layer and a softMax layer as a final layer to analyze the sentiment. Many experiments were conducted with different parameters in the model. The best result achieved was 82% accuracy and 81.6% F-Score with 50 for the LSTM output and 256 for the batch size.

Ombabi et al. [22] built a combined model to classify the polarity of the text in different topics, composed of a CNN as one layer and LSTM as two layers and a support vector machine as a classifier at the final stage. They used Fasttext to represent the words as vectors. By conducting several experiments, the results showed that the proposed model has excellent performance, with an accuracy of up to 90.75%.

Elfaik H. and Nfaoui E. [23] proposed the BILSTM model to classify the sentiment of six Arabic datasets, by encapsulating contextual information from Arabic feature sequences using Forward-Backward encoding. Multiple experiments were conducted on the six datasets, and the results showed that the model obtained excellent results on the Main-AHS dataset with accuracy that reached 92.61%.

Alharbi et al. [24] employed two kinds of recurrent neural networks to build a deep model for Arabic sentiment analysis, LSTM and Gated recurrent units (GRUs) networks. For word vectors, they used FastText vectors, which are considered as input to the networks. Finally, the final output of both networks was classified by three machine learning classifiers. Experiments were conducted on six datasets, and the results of the HTL dataset exceeded the others, they had an accuracy of 94.32%.

3. Arabic Language Characteristics and Challenges

There are three versions of the Arabic language; Classical Arabic (CA) is the language used in the Qur'an, MSA is almost the same as CA and is used and understood in everyday spoken contexts by Arabic speakers around the globe, and the third is informal Arabic (colloquial), which is unstructured in nature . In informal Arabic the individuals speak to one another in a casual manner, which varies from place to place, and the majority of Arabic written on the internet is colloquial Arabic. The Arabic alphabet, in general, comprises 28 letters, except the hamza, with only three of them being vowels. Unlike most other languages, Arabic is written in script, right to left. Also, the shape of the Arabic letters changes by changing their position in the word. A single word, which can be a noun or a verb, in Arabic frequently has many meanings, depending on the context of the sentence. So, Arabic is complex with morphological features language, where many affixes and suffixes are attached to words that change their meaning in a variety of ways. In addition to the process of placing vowels, (fatha:َ, dammah:ُ, kasrah:ِ), above and below the Arabic letters, which is known as diacritization, e.g., the undiacritized word عذب E*b can be interpreted as عَذَّبَ Ea*~aba 'he tortured' and and عَذب Ea*obo 'sweet'. Which is not used in informal Arabic, leading to ambiguity in the interpretation of the intended meaning of words [11], [25].

4. Sentiment Analysis Model

Most studies adopt a model that analyzes sentiments in texts, and it consists of several stages, illustrated in Figure 1) below:

Figure 1: The Sentiment analysis Main Steps

4.1 Datasets for Sentiment analysis

It was demonstrated via what was briefly covered in the second section that the studies conducted to analyze the feelings of the Arabic language do not depend on standard datasets. Most of the datasets were reviews about hotels, books, movies, products, and restaurants. Despite that, the collection of Arabic reviews was not an easy task compared to those written in English, where they were few [26]. Other datasets were collected from social networking sites, with Twitter being the first source among them for the collection task since it offers a search service for the users to search for specific tweets and download them with an API service but with some restrictions [10].

In this section, we will review the datasets that were used to classify sentiment in Arabic texts.

The first dataset used, was the Linguistic Data Consortium Arabic Tree Bank (LDC ATB) dataset, which was in MSA and used by [5] to evaluate their model.

The second one was a Large-Scale Arabic Book Reviews Dataset (LABR), which contains over 63000 book reviews collected and download from Goodreads Site, in two forms MSA and colloquial dialect, submitted by Aly M. and Atiya A. [26]. They investigated it for polarity classification of sentiment, as positive and negative, and rating classification, and it is divided into testing and training. It was used by [7], [20], [22], [23] and [24].

Arabic Sentiment Tweets Dataset (ASTD), presented by Nabil et al. [27], collected from Twitter. There are 10,006 tweets in it, which are divided into four categories: objective, subjective positive, subjective negative, and subjective mixed. It was used by [7], [12], [14], [16], [17], [22], [23], and [24].

Arabic Gold-Standard Twitter Sentiment Dataset that was introduced by Rieser V. and Refaee E. [28], consists of 8,868 tweets collected from Twitter in 2014. It is labelled polar, positive, negative, neutral, and mix. Due to the full dataset not being accessible for free, part of it was used in [7], so they had less accuracy than [28], and also in [8].

Arabic dataset, collected from Twitter, for sentiment analysis, by Abdulla et al. [29], contains 1000 positive tweets and 1000 negative tweets, in a total of 2000 tweets on different topics. Written in the Jordanian dialect and MSA, called ArTwitter. It was used by [7], [14], [22], [23], and [24].

Another dataset introduced by Elsahar et al. [30], included 33k Arabic reviews about different domains collected from websites like TripAdvisor, Qaym, Elcinemas, and Souq, using web crawlers. The dataset contains: Movie Reviews (MOV) consisting of 1.5K reviews, Restaurant Reviews (RES) consisting of 10.9K reviews, Hotel Reviews (HTL) consisting of 15K reviews, and product reviews (PROD) consisting of 15K reviews. The HTL dataset was used by [20]. Also, the HTL and the RES and the PROD were used by the [24]. Moreover, all four categories were used by [26], and were used, in addition to Attraction reviews, which were scrapped from TripAdvisor, by [10].

The authors in [10] collected the dataset from Twitter that was used to test their proposed model, according to two separate eras of time, for two years (2014 and 2015), and (2012 and 2013). It contained 8635 tweets about the economic domain (Egyptian stock exchange), including 2855 positive and 5780 negative, and 3139 tweets about the sports domain, including 407 positive and 2732 negative.

An 1177 random online Al-Jazeera article comments corpus was selected from the QALB dataset [31] by using topic modelling [32], [33], developed by Farra et al. [34], and used by [8] as a test corpus for their model. Also, they were used by [11] to generate the ARSENTB dataset which was used to evaluate their model.

In [16], the authors introduced a dataset collected from Twitter in 2016, that consists of 2026 Arabic tweets on health services. It was classified as 628 positive tweets and 1398 negative tweets manually by three human commentators. It was called the Main-dataset (Main-AHS), and some machine learning and deep learning algorithms were applied to it. Also, it was used by [23]. The same authors extracted 1732 tweets from it that were agreed upon by commentators to be either negative or positive, divided into 502 positive and 1230 negative, and were named the Sub-dataset (Sub-AHS). Both datasets were used by [15], and [16].

The authors in [19] presented a balanced corpus consisting of 20k positive and 20k negative Arabic tweets. The 40k tweets were collected from Twitter on different topics in 2015, written in Egyptian dialect and MSA, and used to conduct their experiments.

4.2 Pre-processing

Pre-processing is an essential phase in many tasks, like the ones with image or text processing, where the data in which imperfections have not been removed leads to misleading results and maybe erroneous after processing. Moreover, the datasets that have been collected from social media and websites were unstructured and often with irrelevant, unreliable, redundant, and noisy data. So, it needs to be clean and clear before inserting it into the model to achieve the desired or expected results.

The text preprocessing phase includes some steps that often should not be sequential, but are determined by the nature of the language, what the analysis is for and from where that data was collected. The main steps are [35]:

- Data cleaning: is considered an important step, where the special symbols, digits, and punctuation marks were removed. For the datasets collected from Twitter, URLs, hashtags symbol (#), mention symbol (@), re-tweet (RT), and usernames were removed. In addition to the above, normalization is used, which is replacing some forms of letters or words with other forms. For the Arabic language, it is applied to letters such as Alif, Ya, Hamzaa and Teh Marbuta. Moreover, Tatweel, diacritics, and letters repeated in the elongation word are normalized by omitting them, [7], [10], [13], [15], [20], [21].
- Stop words removal: It entails removing the most frequently used words from the text that have no effect or importance. Even though occasionally deleting them has minor impact on the outcome. The stop words were removed from the datasets that were used in [10], [20], [21], [23] and [24] Also, a list containing 202 Arabic stop words in [23] was used, as well as a list containing 179 in both [10], [20].
- Tokenization: is the process that involves breaking down a large block of text into smaller tokens. In this scenario, tokens can be words, characters, or subwords. So, it may therefore be divided into three categories: word, character, and subword (n-gram characters) tokenization.
- Stemming: is the process of reducing a word to its most basic form, where prefixes and suffixes at the beginning and end of words are eliminated. The word may be in several inflected forms in the same text or document, which gives the NLP process more redundancy. NLTK ISRI stemmer [36] was the most used stemmer for Arabic texts, where it was used by [9] and [23].
- Lemmatization: is an algorithmic procedure that removes only inflectional ends and returns the base or dictionary form of a word, known as the lemma. It is similar to stemming, but it involves a more thorough series of procedures that include a morphological examination of each word.

What has been observed from previous studies is that there are a few of them that dealt with the emoticons that are present in tweets and reviews [8], [12], [17]. Where emoticons, hashtags, and repeated letters indicate information on word strength and subjectivity of opinion, and can aid in the study of sentiments, and omitting them may lead to the loss of part of the information [35].

4.3 Feature Extraction

After preprocessing phase is completed then the feature extraction process comes, which represented the text in a format that is readable by machines by replacing them with numbers, i.e., a numerical representation of a word.

Word embedding is a feature learning technique in which the contextual hierarchy of words is used to map words to vectors, where the meanings of the word and semantic relationships are captured in it. Feature vectors for comparable words will be the same.

Feature vectors were either created by building models using libraries, with different techniques or different vector dimensions, or by using pre-trained models.
Studies [7], [11], [12], [13], [14], [20], [24] modeled word embedding using word2vec tool [37], [38], training with a CBOW or SG techniques, Doc2vec [39], or Fasttext [40]. [20], [21], [23] employed the Keras library to build word vectors. Other studies like [17], [18], [19], [22] used pre-trained models, such as AraVec [41], Fasttext [42], Glove [43] or word2vec.

4.4 Deep Learning Algorithms

The embedded sentence is passed to the neural network layers once the word embedding layer is applied. The neural networks that were used most for Arabic sentiment analysis were: Deep Neural Network (DNN), which is an artificial neural network (ANN) with multiple layers between the input and output layers that were used with changes in the neurons' number in each layer to achieve the best accuracy for a selected dataset [5], [10], [13]. A convolution neural network (CNN) is a fully connected network that consists of various layers, such as the input layer, convolutional layer, pooling layer, and fully connected layer. Local features within a multi-dimensional field can be recognized using CNNs, by feeding the convolution layer with multidimensional data (such as word embeddings) that consists of numerous filters which in turn learn different features. The work of Kim [44] is the one that used CNN model most for sentence-level sentiment analysis in the English language, which showed it outperformed the traditional machine learning approaches.

A similar model has been adopted in [7], [13]–[15], [17], [19]. RNNs are a type of artificial neural network that has considerable benefits in language modeling. In it, the output is determined not only by the current input but also by the output derived from the previously concealed state of the network. By repeatedly processing each word in a phrase, RNNs store the internal states of the inputs. RNN will thus store all previous words and their relationships in order so that the next word in a phrase will be predicted. Even yet, basic RNN has some issues with information distribution in a long sequence. It was proposed in 1977 that this problem be solved using LSTM, which is an RNN with extra memory [45] that consists of three gates, input, forget and output. used in [14], [17], [19], [21].
Moreover, in 2014 GRU model was presented by [46], to solve the problem of RNN. GRUs are similar to LSTM but do not include the output gate, they work with two types of gates: update and reset, used by [24]. A combination of the LSTM and CNN models was first proposed by [47] to classify sentiment on English tweets. It was also used to analyze sentiments in Arabic texts by [14], [16], [17], [20], and [22]. Direct feedforward neural networks have a limited ability to consider contextual information and hence perform badly in the ASA challenge, therefore, a Bidirectional LSTM model has been proposed. In comparison to LSTM, it learns more semantic information and completely leverages contextual information, as it deals with dependencies in both directions and has the ability to explore more semantic information features from word embedding [48], [49]. It was employed to enhance the Arabic sentiment classification in [23], and, in [18] where it combines with two CNNs.

5. Brief Comparison of Arabic Sentiment Analysis Studies

This section presents a summary and brief comparison between the studies reviewed in this paper in terms of data sets, processing methods, feature extraction methods, classification

models, as well as the highest accuracy obtained by the model. As illustrated in Table 1, note that the values of the fields with the highest accuracy are written in bold.

Table 1: Comparison of Arabic Sentiment Analysis Studies.

Researchers and year	Datasets	Dialect/MSA	Preprocessing	Text representation	Deep learning model	accuracy
Al Sallab et al. (2015) [5]	LDC ATB	MSA	no stemming or removing of stop words.	The Bag of Words, ArSenaL from sentiment lexicon, **Raw words.**	DNN, DBN, DAE, and **Recursive Auto Encoder (RAE).**	74.3%
Dahou et al. (2016) [7]	LABR, ASTD, Gold-Standard, ArTwitter, **HTL**, **ATT**, RES, MOV, PROD.	Dialect and MSA	Removing punctuation, diacritics, and non-Arabic letters, normalize Alef and Teh Marbuta.	word embedding (CBOW and SG)	CNN	For Unbalanced HTL is 91.7%, for unbalanced ATT is 96.2%.
Al-Sallab et al. (2017) [8]	**ATB**, Tweets, online comments from QALB	Dialect and MSA	Removing non-Arabic words, normalization.	word embedding	RAE	86.5%
Abdelhade et al. (2017) [10]	Twitter datasets Egyptian stock exchange, **football tweets**	Dialect	Tokenization, removing non-Arabic letters, punctuation, digits, single Arabic letters, special symbols, username, picture, URL, repeated tweets, Stop Words, and repeated characters, Normalizing Alef, Ya and The Marbuta.	sentiment lexicon	DNN	94.12%
Baly et al. (2017) [11]	ARSENTB	Dialect	Tokenization, Normalization, **stemming**, lemmatization.	word2vec (CBOW)	RNTN	81%
Baly et al. (2017) [12]	ASTD	Dialect	Replace user mention and URL with 'global' tokens, normalize emoticons, , Removing letters repetitions, a	word2vec	RNTN	58.5%

			hashtag symbol and underscore.			
Alayba et al. (2017) [13]	Main-AHS	Dialect and MSA	Removing user mention, URLs, hashtags, Punctuation, diacritics, and Tatweel, Normalizing Alef, The Marbuta, Hamzaa, and letter repetitions.	word2vec	CNN	90%
Al-Azani and El-Alfy (2017) [14]	ASTD, ArTwitter	Dialect and MSA	removing non-Arabic symbols, dialectical marks, punctuation marks, Tatweel, and duplicate character.	word2vec (CBOW and SG)	CNN, LSTM, CNN-LSTM, Stacked LSTM, **combined LSTM**	For ASTD is 81.63%, for ArTwitter 87.27 %
Alayba et al. (2018) [15]	Main-AHS, Sub-AHS	Dialect and MSA	Removed any none Arabic words, digits, and special characters, normalizing Alef and Teh Marbuta.	Word2Vec	CNN	For Main-AHS is 92%, for Sub-AHS is 95%
Alayba et al. (2018) [16]	Main-AHS**, Sub-AHS**, ArTwitter, ASTD	Dialect and MSA	-	word embedding	CNN + LSTM	95.68%
Heikal et al. (2018) [17]	ASTD	Dialect and MSA	Removing digits, re-tweet and mention symbols, diacritics, punctuation, and letters repetitions.	word2vec (SG)	CNN, LSTM, **CNN + LSTM**	65.05%
Wint et al. (2018) [18]	product review, **Social Network Sites**	Dialect and MSA	-	**Word2Vec**(SG), GloVe, fastText	**CNN** + BLSTM	94.38%
Mohammed and Kora (2019) [19]	Tweets Arabic corpus	Dialect and MSA	-	word embedding (CBOW)	CNN, LSTM, **LSTM + augmentation**, RCNN	88.05%
Elzayady et al. (2020) [20]	HTL, **LABR**	Dialect and MSA	Removing non-Arabic numbers, letters, single	Keras' Word Embedding	LSTM, CNN, **combined (CNN-LSTM)**	Average acc 86.88%

			Arabic letters, punctuation, special characters, and letters repetitions.			
Albayati et al. (2020) [21]	LABR	Dialect and MSA	Removing none Arabic words, diacritics, punctuation, letters repetitions, and stop words, Normalizing and stemming.	Keras' Word Embedding	LSTM	82%
Ombabi et al. (2020) [22]	**LABR**, ASTD, ArTwitter	Dialect and MSA	-	Word2Vec, **FastText** (SG and CBOW), AraVec	Deep CNN–LSTM Arabic-SA	90.75%
Elfaik and El Habib (2020) [23]	ASTD, ArTwitter, LABR, MPQA, Multi-Domain, **Main-AHS.**	Dialect and MSA	Tokenization, removing stop words, punctuation, Latin Characters, and digits, Normalizing and stemming.	word embedding	BiLSTM	92.61%
Alharbi et al. (2021) [24]	LABR, ASTD, ArTwitter, **HTL**, RES, PROD.	Dialect	-	Word2Vec (CBOW and SG), Doc2Vec (PV-DBOW AND PV-DM), fastText (CBOW and SG).	DeepASA (GRU, LSTM)	**94.32%**

6. Conclusion and Future Work

The results of experiments employing deep learning to analyze sentiment in Arabic texts revealed that most deep learning models outperformed many of the machine learning models that were examined, including CNN, LSTM. The larger the datasets, the better the model will be trained, and the better results will be obtained. The Arabic language is rich with multi-dialects and needs to be handled carefully to get good results. The review showed that there were small and few datasets that were collected, and most of the preprocessing steps were not standardized, such as normalization and stemming. Moreover, most of the pre-trained feature extraction models were oriented to MSA, while most of the data sets were written in the colloquial dialect of the users in addition to the difference in their writing from one user to another and from one dialect to another. This leads to a weakness in obtaining the important features in the classification.

To improve sentiment analysis in the Arabic language, which is one of the most important and most widely used languages in the virtual world, we suggest that the study be conducted on carefully collected and large data sets. In addition, attention must be put on the dialects of

the Arabic language because it is difficult to obtain large data sets in Standard Arabic only, because most users on the Internet they write in their colloquial language most of the time.

References

[1] L. Zhan, Y. Sun, N. Wang, and X. Zhang, "Understanding the influence of social media on people's life satisfaction through two competing explanatory mechanisms," *Aslib Journal of Information Management*, vol. 68, no. 3, pp. 347–361, May 2016, doi: 10.1108/AJIM-12-2015-0195.

[2] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective Computing and Sentiment Analysis," in *A practical guide to sentiment analysis*, 1st ed., Springer Cham, pp. 1–10, 2017, doi: doi.org/10.1007/978-3-319-55394-8.

[3] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[4] S. AlOtaibi and M. B. Khan, "Sentiment analysis challenges of informal Arabic language," International Journal of Advanced Computer Science and Applications, vol. 8, no. 2, 2017.

[5] A. al Sallab, H. Hajj, G. Badaro, R. Baly, W. el Hajj, and K. Bashir Shaban, "Deep Learning Models for Sentiment Analysis in Arabic," in *Proceedings of the second workshop on Arabic natural language processing*, pp. 9–17, 2015.

[6] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, "A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining," in *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*, pp. 165–173, 2014.

[7] 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*, pp. 2418–2427, 2016.

[8] 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 Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 16, no. 4, pp. 1–20, Jul. 2017, doi: 10.1145/3086575.

[9] R. Baly, G. Badaro, A. Hamdi, R. Moukalled, R. Aoun, G. El-Khoury, A. El-Sallab, H. Hajj, N. Habash, K. B. Shaban, and W. El-Hajj, "OMAM at SemEval-2017 Task 4: Evaluation of English State-of-the-Art Sentiment Analysis Models for Arabic and a New Topic-based Model," in *Proceedings of the 11th international workshop on semantic evaluation (SEMEVAL-2017)*, pp. 603–610, 2017.

[10] 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 *Advances in Intelligent Systems and Computing*, vol. 639, pp. 232–246, 2018, doi: 10.1007/978-3-319-64861-3_22.

[11] 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 Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 16, no. 4, pp. 1–21, 2017.

[12] R. Baly, G. Badaro, G. El-Khoury, R. Moukalled, R. Aoun, H. Hajj, W. El-Hajj, N. Habash, and K. B. Shaban, "A Characterization Study of Arabic Twitter Data with a Benchmarking for State-of-the-Art Opinion Mining Models," in *Proceedings of the third Arabic natural language processing workshop*, pp. 110–118, 2017.

[13] A. M. Alayba, v. Palade, M. England, and R. Iqbal, "Arabic Language Sentiment Analysis on Health Services," *IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pp. 114–118, 2017.

[14] S. Al-Azani and E. S. M. El-Alfy, "Hybrid Deep Learning for Sentiment Polarity Determination of Arabic Microblogs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10635 LNCS, pp. 491–500, 2017, doi: 10.1007/978-3-319-70096-0_51.

[15] 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)*, pp. 13–18, 2018.

[16] 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*, vol. 11015 LNCS, pp. 179–191, 2018, doi: 10.1007/978-3-319-99740-7_12.

[17] 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.

[18] Z. Z. Wint, Y. Manabe, and M. Aritsugi, "Deep learning based sentiment classification in social network services datasets," in *2018 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, pp. 91–96, Nov. 2018, doi: 10.1109/BCD2018.2018.00022.

[19] A. Mohammed and R. Kora, "Deep learning approaches for Arabic sentiment analysis," *Soc Netw Anal Min*, vol. 9, no. 1, pp. 1–12, Dec. 2019, doi: 10.1007/s13278-019-0596-4.

[20] H. Elzayady, K. M. Badran, and G. I. Salama, "Arabic Opinion Mining Using Combined CNN - LSTM Models," *International Journal of Intelligent Systems and Applications*, vol. 12, no. 4, pp. 25–36, Aug. 2020, doi: 10.5815/ijisa.2020.04.03.

[21] A. Q. Al-Bayati, A. S. Al-Araji, and S. H. Ameen, "Arabic Sentiment Analysis (ASA) Using Deep Learning Approach," *Journal of Engineering*, vol. 26, no. 6, pp. 85–93, Jun. 2020, doi: 10.31026/j.eng.2020.06.07.

[22] A. H. Ombabi, W. Ouarda, and A. M. Alimi, "Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks," *Soc Netw Anal Min*, vol. 10, no. 1, Dec. 2020, doi: 10.1007/s13278-020-00668-1.

[23] H. Elfaik and E. H. Nfaoui, "Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 395–412, Jan. 2021, doi: 10.1515/jisys-2020-0021.

[24] A. Alharbi, M. Kalkatawi, and M. Taileb, "Arabic Sentiment Analysis Using Deep Learning and Ensemble Methods," *Arab J Sci Eng*, vol. 46, no. 9, pp. 8913–8923, Sep. 2021, doi: 10.1007/s13369-021-05475-0.

[25] S. Alotaibi and M. B. Khan, "Sentiment Analysis Challenges of Informal Arabic Language," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 2, pp. 278–284, 2017.

[26] 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)*, pp. 494–498, 2013.

[27] M. Nabil, M. Aly, and A. F. Atiya, "ASTD: Arabic Sentiment Tweets Dataset," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2515–2519, 2015.

[28] V. Rieser and E. Refaee, "An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis," in *LREC*, pp. 2268–2273, 2014.

[29] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and Al-Ayyoub M., "Arabic sentiment analysis: Lexicon based and corpus-based," in *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pp. 1–6, 2013.

[30] 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*, pp. 23–34, 2011.

[31] 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)*, pp. 39–47, 2014.

[32] Blei D. M., Ng A. Y., and Jordan M. I., "Latent Dirichlet Allocation Michael," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[33] McCallum and A. Kachites, "MALLET:A Machine Learning for Language Toolkit," *http://mallet.cs.umass.edu*, 2002.

[34] N. Farra, K. Mckeown, and N. Habash, "Annotating Targets of Opinions in Arabic using Crowdsourcing," in *Proceedings of the second workshop on Arabic natural language processing*, pp. 89–98, 2015.

[35] Z. Nassr, N. Sael, and F. Benabbou, "Preprocessing Arabic dialect for sentiment mining: State of art," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 44, pp. 323–330, Nov. 2020.

[36] 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. IEEE*, vol. 1, pp. 152–157, 2005.

[37] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in neural information processing systems*, vol. 26, pp. 3111–3119, 2013.

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proceedings of Workshop at International Conference on Learning Representations*, 2013.

[39] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *International conference on machine learning, PMLR*, pp. 1188–1196, 2014.

[40] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Trans Assoc Comput Linguist*, vol. 5, pp. 135–146, 2017.

[41] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," *Procedia Comput Sci*, vol. 117, pp. 256–265, 2017, doi: 10.1016/j.procs.2017.10.117.

[42] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in Pre-Training Distributed Word Representations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.

[43] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[44] Y. Kim, "Convolutional neural networks for sentence classification.," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735-1780, 1997.

[46] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014, doi: 10.3115/v1/D14-1179.

[47] P. M. Sosa and C. Yang, "Twitter Sentiment Analysis using combined LSTM-CNN Models," *Eprint Arxiv*, pp. 1–9, 2017.

[48] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for Target-Dependent Sentiment Classification," in *International Conference on Computational Linguistics*, pp. 3298–3307, Dec. 2016.

[49] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "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)*, pp. 207–212, 2016.