Arabic Sentiment Analysis Using Deep Learning: A Review

Zainab Hakami [1†], Muneera Alshathri, Nora Alqhtani, Latifah Alharthi and Sarah Alhumoud [2††]
Umm AlQura University, College of Computer and Information systems, Makkah, Saudi Arabia 1†
Al Imam Mohammad Ibn Saud Islamic University, (IMSIU), Saudi Arabia 2††

Summary
Social media provides a significant source of public opinions and trends. Recently, the interest in analyzing this publicly available data through sentiment analysis has increased noticeably. The use of deep-learning for sentiment analysis is lately under focus, as it provides a scalable and direct way to analyze text without the need to manually feature-engineer the data. As the work on Arabic sentiment analysis using deep learning is scarce and scattered, this paper presents a systematic review of those studies covering the whole literature, analyzing 19 papers. The review proves a general trend of Arabic sentiment analysis performance improvement with deep learning as opposed to sentiment analysis using machine learning.

Key words:
Arabic, sentiment analysis, deep learning, CNN, RNN, NN.

1. Introduction

Artificial intelligence has contributed in saving humans' effort and time in doing tedious tasks and also enhanced performance efficiency and throughput in multiple domains. Artificial intelligence (AI) is "the study of the computations that make it possible to perceive, reason, and act" [1].

Natural language processing (NLP) is a combination field of computational linguistics and artificial intelligence that aims to achieve human-like language processing [2]. Unlike other data processing fields in AI, an NLP system requires knowledge of the language. This knowledge consists of different levels and types, for example, phonetics and phonology, morphology, syntax, semantics, pragmatics, and discourse [2]. Additionally, NLP can apply tasks through several approaches such as the machine learning approach.

Machine learning (ML) is a branch of AI that aims to find the meaning of data patterns through computers [3]. ML aims to make computers learn from an input. The training data in learning is the input, which is an experience, and the output is expertise [3]. Machine learning is not about attempting to construct intelligent behavior based on automated imitation. It executes tasks that humans can perform. ML could be classified into two broad categories—supervised and unsupervised learning. Supervised learning requires training on labeled data. Unsupervised learning clusters the inputs into subsets of similar features [3].

Deep learning is a specific type of machine learning that processes information through a neurons structure inspired by the human brain; it extracts features from data and has been recently applied in fields such as image and video recognition/classification, audio processing, text analysis and NLP tasks, autonomous systems and robotics, medical diagnostics, computational biology, physical sciences, finance and market analysis, cyber security, and algorithmic enhancement [4]. Deep learning could be classified into supervised, unsupervised, and reinforcement learning, which is a combination of the former two by assigning a reword to each correct prediction [4].

Sentiment analysis (SA) is the field of study that depends on natural language processing [5]. SA aims to analyze opinions with emotions and classify them to be positive or negative sentiments. SA is also known as subjectivity analysis, emotion analysis, opinion mining, opinion extraction, sentiment mining, effect analysis, and review mining [6]. Additionally, SA is mainly implemented through supervised machine learning or unsupervised lexicon-based methods using approaches initially proposed by Pang et al. [7] and Liu [6].

One of the challenges of SA using machine learning is the requirement of a large labeled dataset for the model to be trained on and learn from. This is alleviated with the use of deep learning approaches for SA that does not depend on extensive manual feature engineering by extracting the features automatically [8]. This is an advantage over other SA approaches, which are more robust and accessible to generalize to other domains with better performance and dimensionality reduction [8], [9], [5], [6]. Additionally, the research on Arabic NLP, in general, is challenging due to the characteristics of the Arabic language in "morphological richness, ambiguity, and lexical sparsity" [10]. Furthermore, having multiple forms of Arabic language—classical, standard, and dialectal—poses another challenge [11]. Dialects are used mainly in informal spoken communication and is now used extensively in written communication in social media with multiple forms that differ in each region.

The rest of this paper is organized as follows: in Section II, the review methodology includes the details of this paper's review including searched databases, and in Section III, the

Manuscript received April 5, 2019
Manuscript revised April 20, 2019

related work are presented as the literature review. Finally, Section IV concludes the paper.

2. Review Methodology

This review is inspired by the methodology described in [28], [29]. Additionally, the focus of this review is the literature available on Arabic Sentiment Analysis (ASA) using deep learning during the period 2015, when the first paper on the subject was published, to 2018. The focal points are:

- Arabic Neural Networks (ANN)
- Arabic Convolutional Neural Networks (ACNN)
- Arabic Recursive Neural Networks (ARNN)

The databases we used for our research are IEEE, Sage, ACM, ACL web (Association for Computational Linguistics), Springer, and Science Direct. The keywords used are: “x y”, where x is “Arabic Sentiment” and y is either” neural,” “convolutional,” or “recurrent.”

After eliminating papers that are out of the focus of this review, we were left with 19 papers that matched our criteria. Figure 1 shows the distribution of the papers across the different databases.

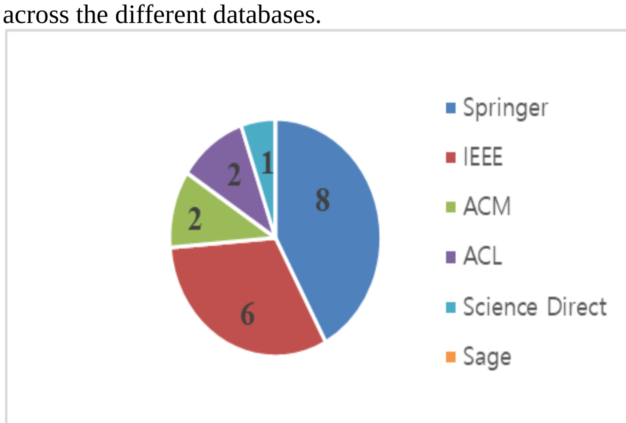

Figure 1: papers found in each database

3. Literature Review

There are three main steps that have been applied to investigate the literature. Arabic sentiment analysis: (1) gathering texts, (2) pre-processing and annotating text polarity (either positive or negative) on the dataset (unless mentioned otherwise), then (3) applying deep learning. In Arabic texts, pre-processing is done by removing non-Arabic symbols, removing diacritics (harakat), removing punctuation marks, removing the stretching character (tatweel or kashida), and removing duplicate characters.

Generally, there are two well-known deep learning techniques that are used for Arabic sentiment analysis. Those are recurrent neural networks (RNN) and convolutional neural networks (CNN) [8], and each has been discussed in a separate category in our review. Moreover, we found that some papers used other basic deep learning techniques such as deep neural network (DNN), deep belief networks (DBN), deep auto encoder (DAE), and recursive auto encoder (RAE), and these were collected in the NN category.

3.1 Arabic Neural Networks (ANN)

To insert A neural network is a collection of interconnected processing elements or units that are organized in three interconnected layers: input, output, and hidden layers. The functionality of the network is inspired by brain neurons, that is, it learns and performs tasks such as predictions and classification [38]. In the context of language processing, neural networks aim to learn the syntactic and semantic representations of language in the real-valued continuous space in a way by which similar words or structures are near each other [39]. We have reviewed the available literature on Arabic sentiment analysis using neural networks in the following paragraphs. This review covers works that apply NN in different ways with a total of 5 papers having been published by 4 different authors; most of these papers applied DNN and the highest accuracy score was for Abdelhade with the value of 90.22%.

Al Sallab et al. [12] explored four deep learning models: deep neural networks (DNN), DBN, DAE, and RAE. They used the Arabic Tree Bank (ATB) dataset which contains 1180 sentences, with a bag of a word (BOW) for word representation. In the first three models (DNN, DBN, and DAE), they employed features based on ArSenL in both lemma and sentences as well as raw words. The accuracy for DNN scored 55.5% and 53.4% for lemma and sentences, respectively, and 39.5 in raw words. DAE scored 60.4% for lemma, 56.1% for sentences, and 43.5% for raw words. DBN scored 57.5% for lemma, 53.4% for sentences, and 41.3% for raw words. The fourth model RAE scored the best result with 74.3% in accuracy for raw words.

Moreover, Al-Sallab et al. [13] presented a model named Recursive Deep Learning Model for Opinion Mining in Arabic (AROMA) which optimize some options such as: adding morphological tokenization, enhance embeddings initialization by combining vectors of both semantic and sentiment metrics and leveraging BOW representation. AROMA was examined on three Arabic text datasets: (1) ATB with corpus size of 1180, (2) Qatar Arabic Language Bank (QALB) with corpus size of 1133, and (3) tweets with corpus size of 2311 tweets. The framework was compared to the baseline RAE model proposed in [12], The accuracy of AROMA outperforms the baseline in all the three datasets by values 86.5%, 79.2%, and 76.9% on ATB, QALB, and Tweets, respectively. Even though AROMA discussed bright optimization details, the small dataset used limited the expected enhancement.

Abdelhade et al. [14] applied ASA using DNN classifier with backpropagation algorithm. They analyzed Arabic tweets in two different domains: Egyptian stock exchange and sports tweets. For annotating the dataset, they constructed a lexicon, NileULex [15], from two already existing ones and a fully automated approach for Arabic slang lexicon extraction from microblogs [16]. The resulting lexicon contains 15840 unique words with 9595 positives and 6245 negatives. A sentiment lexicon is a numerical representation for the semantic of each unique word, the numbers of all words in the text are summed to detect overall polarity of given text. The experiments are conducted using three classifiers NB, Decision Tree (D-Tree), and K-Nearest Neighbour (KNN) to compare them with the performance of DNN. The experiments show that the DNN outperforms other three machine learning algorithms evaluated with an average accuracy of 90.22%, precision 90.56%, recall 90.90%, and F-measure of 90.68%.

Ahmad et al. [37] applied a sentiment classification for a set of 11 deep emotions. Some to mention, anticipation, optimism, and pessimism. DNN was used in two different ways, single DNN, and multi DNN. Where single DNN refers to adding features as nodes within the classifier training the output layer. While multi DNN points to training the classifier on one feature at a time which creates multiple DNNs. Two datasets extracted from SemEval-2018 were used to test the framework, one was 7686 English tweets, where the other was 2785 Arabic tweets. The single DNN scored 62.4% and 60.1% in accuracy for the English and Arabic datasets respectively. Whereas the multi DNN scored 62.4% and 61.1% for both datasets which implies no significant differences in the accuracy for both single and multi DNN.

Essatouti et al. [17] proposed a new text data representation model based on BOW by adding distance features computed using the Levenshtein Distance. Then, they applied DNN on a dataset made of Moroccan dialect comments to detect their polarity. The dataset consists of 1490 comments from different political articles from Moroccan Hespress news website. The collected comments are mainly in Modern Standard Arabic (MSA) or in the Moroccan dialectal Arabic; some of the comments are a mix between the two. Furthermore, they achieved 62% accuracy by the distance-based method. It is clear that they have a small data set that cannot exploit deep learning advantages.

3.2 Arabic Convolutional Neural Networks (ACNN)

Convolutional Neural Network (CNN) is a type of artificial neural networks that is similar to the definition presented in the previous section, and it has input and output layers in addition to multiple hidden layers. The difference here is that some of these layers are convolutional, using a mathematical model to pass on weighted results to successive layers [40]. Furthermore, CNN is used in multiple domains such as computer vision and NLP. In the following paragraphs, we have reviewed the available literature on Arabic sentiment analysis using CNN. This review covers a total of 7 papers published by 5 different authors. Alayba outperforms others in terms of the number of papers as well as in gaining the highest scored accuracy of 94.24%.

Dahou et al. [18] created a huge web-crawled corpus of about 10 billion words from two different domains—tweets and reviews—to apply a binary sentiment classification. They used the CBOW and Skip-Gram models as word representations. They applied CNN in two ways: CNN-balanced, where an equal number of positive and negative texts is considered in the dataset, and, otherwise, CNN-unbalanced. They used the LABR dataset that contains more than 63,000 book reviews. The accuracy score was 89.6% for unbalanced and 86.7% for balanced. However, the Arabic sentiment tweets dataset (ASTD) includes over 10,000 Arabic tweets with an accuracy of 79.07% for unbalanced and 75.9% for balanced. Whereas, the Arabic gold-standard twitter sentiment corpus that was collected by Refaee and Rieser in 2014, scored 75.8% for unbalanced and 73.8% for balanced. Moreover, the Arabic twitter dataset collected by Abdulla et al. in 2013 [19] consist of 2000 labeled tweets, and scored 85.01% for unbalanced and 86.3% for balanced. The datasets collected by ElSahar and El-Beltagy [20] covers five domains: hotel reviews consist of 15,000 Arabic reviews and scored 91.7% for unbalanced and 88.6% for balanced, attraction reviews scrapped from TripAdvisor scored 96.2% for unbalanced and 74.2% for balanced, restaurant reviews from two resources (Qaym and TripAdvisor) scored 78.5% for unbalanced and 77.1% for balanced, movie reviews consist of 1500 movies reviews and scored 80.7% for unbalanced and 83.2% for balanced, and product reviews consist of 15,000 reviews and scored 87.3% for unbalanced and 83.3% for balanced. As shown above, the best results go to the unbalanced dataset, the size and quality of data also influence performance noticeably.

Al-Azani and El-Alfy [21] investigated various deep learning models based on CNN and LSTM for sentiment analysis of Arabic microblogs. They adopted a neural language model created by Google, known as word2vec, for vectorizing text. Word2vec has two neural network architectures based on different word representations: (CBOW) and skip-gram. The experiments were run on ASTD, which consists of 10,000 tweets, and ArTwitter [20], which holds 2000 Arabic tweets. There are four models of LSTM: Simple LSTM, CNN-LSTM, Stacked LSTM, and Combined LSTM. The experiments show that using word2vec vectors that are updated during learning achieves the highest results in nearly all cases. Additionally, Al-Azani and El-Alfy [21] proved that LSTM performs better than CNN. Moreover, the proposed combined LSTM architectures perform better than the other models with an accuracy of 87.27%. Both datasets, ASTD and ArTwitter, were applied and compared on the different word representation stated earlier on static and non-static word initialization models. Table 1 shows the detailed results of the tested models.

Alayba et al. [34] built an Arabic text dataset on health services tweets. They used and compared the performance of several machine and deep learning algorithms: NB, SVM, logistic regression (LR), DNN and CNN with Word2Vec. The data was collected from Twitter trending hashtags between 1st February 2016 and 31st July 2016. The dataset was named Arabic Health Services Dataset (AHS) and has been manually polarity annotated by three annotators. The reason for having three judges is to get three different opinions about each tweet, then calculating the majority vote. The dataset has 2026 tweets after filtering (removing tweets out of topic), 628 positive tweets and 1398 negative tweets. The accuracy result was 85% with the best classifier SVM; furthermore, machine learning outperforms CNN due to the small size of the dataset.

Alayba et al. [22] continued with their work presented in [34] and, in this paper, they constructed a Word2Vec special model maintained by large Arabic corpus obtained from ten Arabic newspapers from different countries in order to apply word representation on previous Arabic health services dataset with CBOW model. CNN was used integrated with the Lexicon models. Different text features of selection methods using numerous machine learning classifiers have been experimented and explored, and the best result was achieved by SVM with a 7% inaccuracy improvement from their past results in AHS [34]. This result verifies that using CNN integrated with lexicon upgraded the performance of the classification model.

Furthermore, Alayba et al. [23] investigated the benefits of integrating CNN and LSTMs. Furthermore, CNN offers advantages in selecting good features, and LSTM networks have proven good abilities to learn sequential data. Alayba et al. [23] used AHS in [34], and they explored the effectiveness of using different levels of sentiment analysis: character level, character N-gram level and word level. Due to the complexities of morphology and orthography in Arabic, they found that using word-level has shown better sentiment classification result and improved accuracy on AHS reach to 94.24%, as compared to their previous results in [23], which was 92%. The model was compared with other know datasets (i.e., ArTwitter and ASTD) but they got fewer accuracies, 88.10% and 76.41%, respectively. Additionally, the performance improved by combining different deep learning techniques even with a small dataset; this proved the power of deep learning in ASA.

Gridach et al. [25] explored a new method without using any kind of hand-crafted features such as word pre-processing or traditional machine learning approaches. They represented the CNN-ASAWR system as the abbreviation of CNN for ASA using word representations; furthermore, the system was tested by three main word representations: Stanford Glove vectors, Skip-gram model and CBOW. They used ASTD dataset with 10006 tweets and SemEval 2017 with 3355 tweets. The outcomes show that the best F1-score results for CBOW model in both datasets were 72.14% and 63.00%, accordingly. To conclude, although the model performed worst without pre-processing, on the other hand, it saved time and effort with satisfying results, which need more experiments to prove its efficiency.

Abdullah et al. [35] proposed a SEDAT system (Sentiment and Emotion Detection for Arabic Text) that detects the type of emotions (anger, joy, fear, and sadness) in text and its intensity (low, moderate, and high) along with sentiment (neutral, positive, and negative) by analyzing both text and emojis through CNN-LSTM. The datasets used to explore the system was 7400 Arabic tweets from ArTweets after translating it to the English language in order to apply the best preprocessing tools. The results for emotion classification accuracy was 56.9%, while sentiment gained higher accuracy was 78.6%. Additionally, translating texts may lead to a false representation of some words, which lowers the overall performance.

3.3 Arabic Recurrent Neural Networks (ARNN)

A recurrent neural network (RNN) is a type of neural network with closed loop connections, allowing information to persist [41]. This happens when the recurrent net elements have two inputs, the present and the recent past, that aid in better learning. In the following paragraphs, we have reviewed the available literature on Arabic sentiment analysis using RNN. This review covers a total of 7 papers published by different authors' the highest accuracy scored was for Alwehaibi and Roy, with a value of 93.5%.

Abbes et al. [26] presented an experiment where the performance of DNN and RNN were compared, which were applied on the same dataset collected from the LABR book reviews with a total of 1800 review; furthermore, for

word representation they adopted BOW. Comparing the results proved that RNN outperforms DNN in accuracy with a gap of 7.6%. To be precise, DNN accuracy was 64.4% while RNN performance gives 71% correct classification. RNN overcomes DNN due to its ability in creating long-term dependencies through recurrent looping.
El-Kilany et al. [27] addressed the issue of recognizing entities that are targeted by text sentiment in Arabic tweets. The sentiment target recognition can be accomplished by two main tasks: named entity recognition and sentiment analysis. Named entity recognition discovers all named entities in the given text that affect the overall text's sentiment, while sentiment analysis analyzes the text to either positive or negative. The proposed model used was Bidirectional Long Short-Term Memory Networks (BI-LSTMs) together with Conditional Random Fields (CRF) classification layer evolved with a word embeddings layer. The model was evaluated on the annotated Arabic tweets dataset. The dataset selected was 3000 tweets. The precision metric scored 73.7%. The results show that the model attained promising results in the Arabic recognition entity field.
Baly et al. [10] created the Arabic sentiment treebank (ARSENTB) for the first time to apply with the Recursive Neural Tensor Networks (RNTN) model in order to overcome morphological hardships in Arabic. They compared their result with other different classifiers (SVM, RAE, and LSTM) to check for best accuracy; the dataset used 1177 comments on newspaper articles, with CBOW as word representation. The results proved the best values for RNTN in the phrase and comment level with 83% and 80% inaccuracy, respectively. Such a small dataset was not enough to reflect the population representative data for the proposed model.
Al-Smadi et al. [28] compared the performance for the two approaches, RNN and SVM, by adding N-gram as word representation on a dataset of 2291 Arabic hotel reviews. The results revealed that the best accuracy was for SVM with a total of 95.4%, while RNN proved the best speed in execution time but with lower accuracy of 87%; this could be due to the dataset being small and the performance of deep learning being enhanced with the increase of the dataset size.
In addition, Al-Smadi et al. [24] experiment Aspect-Based Sentiment Analysis (ABSA) dataset on Arabic Hotels' reviews. They used two implementations based on LSTM: first, a character-level BI-LSTM combined with CRF for aspect opinion target expressions (OTEs) extraction; second, an Aspect-Based LSTM for aspect sentiment Polarity Classification (AB-LSTM-PC). The dataset contains 24,028 annotated reviews and was annotated manually using the SemEvalABSA16 annotation guidelines on both the text and sentence level. Only the sentence level tasks have been targeted in this research. They use word2vector for word representation. The results show that the two applied implementations are better compared to the SVM results with the accuracy of 82.6%. This proves that deep learning results have surpassed machine learning approaches in ASA.
Al-Azani et al. [36] investigated the polarity of Arabic texts with the aid of emojis extracted features verified by the Emoji Sentiment Ranking (ESR) lexicon. By applying and comparing both LSTM and Gated Recurrent Unit (GRU) on manually annotated dataset consist of 2091 microblogs containing emojis were extracted from Twitter and YouTube. The best results appeared when the two approaches were in an accuracy value of 78.71%.
Alwehaibi and Roy [29] inspected an LSTM-RNN model for the Arabic text sentiment analysis and compared different pre-trained Word Embedding (WE) models that affect the model's accuracy. They used the AraSenTi-Tweet [30] dataset, which is an Arabic corpus collected from Twitter to test text polarity. Notably, they tested three publicly available Arabic pre-trained WEs used to generate word vectors, which are AraVec [31], Arabic FastTxt [32], which is called AraFT hereafter, and Arabic news [33]. The results of their experiment show that LSTM-RNN model achieves the highest accuracy of 93.5% with AraFT, then ArabicNews with 91%, and lastly, AraVec with 88%. AraFT achieved the best because it was built for the purpose of overcoming the morphological complexity of the Arabic language.

4. Conclusion

This paper reviews studies on the Arabic sentiment analysis using deep learning during the period from 2015 to 2018. The literature with this regard is limited compared to the studies on sentiment in the English texts, as only 19 papers were found. The studies were mainly in three different areas: ANN, ACNN, and ARNN. As for the number of studies found in each category, the count was 5 for ANN, and 7 each for both ACNN and ARNN. The results show that using deep learning for sentiment analysis generally increases the model accuracy compared to using machine learning. However, the results also show that when the database is small, the performance of deep learning model is degraded noticeably.

Table 1: Literature summary

Article	***Context***	***Word representation***	***model***	***Dataset***	***Accuracy %***		
[26]	Reviews	BOW	DNN	LABR = 1800 reviews	64.4		
			RNN		71.00		
[23]	Twitter	N-gram Word-level	LSTM CNN	AHS [34] = 2026 tweets	94.24		
				ArTwitter = 2000 tweets	88.10		
				ASTD =10000 tweets	76.41		
[12]	Newswire	BOW	DNN	ATB = 1180 Sentences	**Lemma**	**Sentence**	**Raw words**
					55.5	53.4	39.5
			DAE		60.4	56.1	43.5
			DBN		57.5	53.4	41.3
			RAE		-	-	74.3
[13]	Newswire	BOW	RAE	ATB = 1180 Sentences	86.5		
	Online comments			QALB =1113 comments	79.2		
	Twitter			2311 tweets	76.9		
[14]	Twitter	-	DNN	15840 tweets	90.22		
[17]	News website comments	BOW and DistanceBased	DNN	1490 comments	62.00		
[21]	Twitter	CBOW	CNN	ASTD =10000 tweets	***static***	***Non-static***	
					74.40	74.10	
			LSTM		74.70	80.12	
			CNN-LSTM		68.07	73.49	
			Stacked-LSTM		65.66	70.18	
			Combined LSTM-SUM		78.31	81.02	
			Combined LSTM-MUL		77.41	81.03	
			Combined LSTM-CONC		77.11	80.42	
			CNN	ArTwiter = 2000 tweets	77.21	77.82	
			LSTM		83.16	84.39	
			CNN-LSTM		78.23	80.70	
			Stacked-LSTM		82.34	81.93	
			Combined LSTM-SUM		82.55	84.80	
			Combined LSTM-MUL		82.96	85.42	
			Combined LSTM-CONC		82.96	86.45	
		S-gram	CNN	ASTD =10000 tweets	61.45	66.57	
			LSTM		76.51	77.41	
			CNN-LSTM		75.90	71.99	
			Stacked-LSTM		68.98	76.51	
			Combined LSTM-SUM		78.31	78.92	
			Combined LSTM-MUL		77.11	76.20	
			Combined LSTM-CONC		78.61	80.42	
			CNN	ArTwiter = 2000 tweets	75.56	83.16	
			LSTM		80.90	83.57	
			CNN-LSTM		73.92	84.19	

			Stacked-LSTM		81.72	82.96
			Combined LSTM-SUM		82.55	85.63
			Combined LSTM-MUL		81.72	85.83
			Combined LSTM-CONC		81.31	87.27
[18]	Twitter and Reviews	CBOW S-gram	CNN	LABR = 58,713 review	**Balanced**	**Unbalanced**
					89.6	86.7
				ASTD = 2826 tweets	75.9	79.07
				Arabic Gold-Standard = 3613 tweets	73.8	75.8
				Dataset in [19] = 2000 tweets	86.3	85.01
				Hotel Reviews [20] = 39812 reviews	88.6	91.7
				Attraction Reviews [20] = 5650 reviews	74.2	96.2
				Restaurant Reviews [20] = 44042 reviews	77.1	78.5
				Movie Reviews [20] = 1737 reviews	83.2	80.7
				Product Reviews [20] = 4496 reviews	83.3	87.3
[24]	Reviews	Word2vect	LSTM	ABSA = 24,028 reviews	82.6	
[34]	Twitter	Word2vect	DNN	AHS [34] = 2026 tweets	85.0	
			CNN		90	
[22]	Twitter	Word2vect	CNN	AHS [34] = 2026 tweets	95	
[25]	Twitter	CBOW	CNN	SemEval 2017 = 3355 tweets	63.00	
				ASTD =10000 tweets	72.14	
[27]	Twitter	Word2vect	BI-LSTMs CRF	3000 tweets	73.7	
[10]	News website comments	CBOW	RNTN	1177 comments	**Phrase level** 83	
					Comment level 80	
[28]	Reviews	N-gram	RNN	Arabic hotel review =2291 reviews	87	
[29]	Twitter	AraFT	RNN-LSTM	AraSenti = 10,000 tweets	93.5	
		Arabic News			91	
		AraVec			88	
[37]	Twitter	-	DNN	SemEval-2018 = 7686 English tweets	**Single DNN**	**Multi DNN**
					62.4	62.4
				SemEval-2018 = 2785 Arabic tweets	60.1	61.1
[35]	Twitter	-	CNN-LSTM	ArTwiter = 7400 tweets	**Sentiment** 78.6	
					Emotions 56.9	
[36]	Twitter and YouTube	-	LSTM-GRU	2091 microblogs	78.71	

References

[1] P. Henry, Artificial Intelligence, 3rd ed. USA, Addison-Wesley Publishing Company, 1992.

[2] D. Jurafsky and J. H. Martin, Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd ed. New Jersey, Pearson Prentice Hall, 2008.

[3] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge, UK: Cambridge University Press, 2014.

[4] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," IEEE Access, vol. 6, pp. 24411–24432, 2018.

[5] L. Deng and Y. Liu, Deep Learning in Natural Language Processing. New York, NY: Springer Berlin Heidelberg, 2018.

[6] Bing Liu, Sentiment Analysis and Opinion Mining. Chicago, Morgan & Claypool, 2012.

[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP)- Volume 10, Stroudsburg, PA, USA, 2002, pp. 79–86.

[8] D. Stojanovski, G. Strezoski, G. Madjarov, I. Dimitrovski, and I. Chorbev, "Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages," Multimedia Tools and Applications, Volume 77, Issue 24, pp 32213-32242, 2018.

[9] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for twitter sentiment analysis," IEEE Access, vol. 6, pp. 23253–23260, 2018.

[10] 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, Jul. 2017.

[11] N. Y. Habash, "Introduction to Arabic Natural Language Processing". Morgan & Claypool, vol. 37, pp. 623–625, Sep. 2011.

[12] 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 (WANLP), Beijing, China, 2015, pp. 9–17.

[13] 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 and Low-Resource Language Information Processing (TALLIP), vol. 16, no. 4, pp. 25:1–25:20, Jul. 2017.

[14] 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 (AISI2017), Cairo, Egypt, (vol. 639), pp. 232–246.

[15] S. R. El-Beltagy, "NileULex: A phrase and word level sentiment lexicon for Egyptian and modern standard Arabic," in LREC 2016, Portorož, p. 6.

[16] H. ElSahar and S. R. El-Beltagy, "A fully automated approach for Arabic slang lexicon extraction from microblogs," in Computational Linguistics and Intelligent Text Processing (CICLing), (vol. 8403), A. Gelbukh, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 79–91.

[17] Essatouti, H. Khamar, S. E. Fkihi, R. Faizi, and R. O. H. Thami, "Arabic sentiment analysis using a levenshtein distance based representation approach," in 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), Marrakech, 2018, pp. 270–273.

[18] 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, Osaka, Japan, 2016, pp. 2418–2427.

[19] 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.

[20] Gelbukh and CICLing, Eds., Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14–20, 2015.

[21] S. Al-Azani and E.-S. M. El-Alfy, "Hybrid deep learning for sentiment polarity determination of Arabic microblogs," in Neural Information Processing, 2017, pp. 491–500.

[22] M. Alayba, V. Palade, M. England, and R. Iqbal, "Improving sentiment analysis in Arabic using word representation" in 2nd International Workshop on Arabic Script Analysis and Recognition (ASAR), London, UK, 2018.

[23] M. Alayba, V. Palade, M. England, and R. Iqbal, "A combined CNN and LSTM model for Arabic sentiment analysis," Machine Learning and Knowledge Extraction, vol. 11015, pp. 179–191, 2018.

[24] 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," International Journal of Machine Learning and Cybernetics, vol. 10, issue 55, pp.1-13, 2018.

[25] M. Gridach, H. Haddad, and H. Mulki, "Empirical evaluation of word representations on Arabic sentiment analysis," Arabic Language Processing: From Theory to Practice, vol. 782, pp. 147–158, 2018.

[26] M. Abbes, Z. Kechaou, and A. M. Alimi, "Enhanced deep learning models for sentiment analysis in Arab social media," Neural Information Processing, vol. 10638, pp. 667–676, 2017.

[27] El-Kilany A., Azzam A., El-Beltagy S.R., "Using Deep Neutral Networks for Extracting Sentiment Targets in Arabic Tweets". In Shaalan K., Hassanien A., Tolba F. (eds) Intelligent Natural Language Processing: Trends and applications, SCI, vol 740, 2018, pp. 3-15, Springer, Switzerland, Cham.

[28] 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,” Journal of Computational Science, vol. 27, pp. 386–393, Jul. 2018.
[29] Alwehaibi and K. Roy, “Comparison of pre-trained word vectors for Arabic text classification using deep learning approach,” in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, 2018, pp. 1471–1474.
[30] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, “AraSenTi-Tweet: A corpus for Arabic sentiment analysis of Saudi tweets,” Procedia Computer Science, vol. 117, pp. 63–72, 2017.
[31] B. Soliman, K. Eissa, and S. R. El-Beltagy, “AraVec: A set of Arabic word embedding models for use in Arabic NLP,” Procedia Computer Science, vol. 117, pp. 256–265, 2017.
[32] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, “Learning Word Vectors for 157 Languages,”. arXiv preprint arXiv:1802.06893, 2018
[33] A. Altowayan and L. Tao, “Word embeddings for Arabic sentiment analysis,” in 2016 IEEE International Conference on Big Data (Big Data), Washington DC, USA, 2016, pp. 3820–3825.
[34] M. Alayba, V. Palade, M. England, and R. Iqbal, “Arabic language sentiment analysis on health services,” in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, pp. 114–118.
[35] M. Abdullah, M. Hadzikadicy, and S. Shaikhz, “SEDAT: Sentiment and emotion detection in Arabic text using CNN-LSTM deep learning,” in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, pp. 835–840.
[36] S. Al-Azani and E.-S. El-Alfy, “Emojis-based sentiment classification of Arabic microblogs using deep recurrent neural networks,” in 2018 International Conference on Computing Sciences and Engineering (ICCSE), Kuwait City, pp. 1–6.
[37] T. Ahmad, A. Ramsay, and H. Ahmed, “Explorations in sentiment mining for Arabic and English tweets,” in Artificial Intelligence: Methodology, Systems, and Applications, vol. 11089, G. Agre, J. van Genabith, and T. Declerck, Eds. Switzerland, Cham: Springer International Publishing, 2018, pp. 16–24.
[38] K. Gurney, An Introduction to Neural Networks. London, CRC Press, 1997.
[39] J. Zhang and C. Zong, “Deep neural networks in machine translation: An overview,” IEEE Intelligent Systems, vol. 30, no. 5, pp. 16–25, Sep. 2015.
[40] Goodfellow, Y. Bengio and A. Courville, Deep Learning. .USA, MIT press, 2016.
[41] L. Medsker and L. C. Jain, Recurrent Neural Networks: Design and Applications. USA, CRC Press, 1999.