Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Sentiment Analysis in Arabic Text and Emoji Using Deep Learning Methods

Suhad Nasrallah Taraf

Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics, Baghdad, Iraq
ms202210721@iips.edu.iq

Ahmed T. Sadiq

Department of computer science, College of Science, University of Technology-Iraq, Baghdad, Baghdad, Iraq
drahmaed_tark@yahoo.com

ARTICLEINFO
Article History
Received: 13/10/2024
Accepted: 24/10/2024
Published: 31//2024

ABSTRACT The advent of social media has simplified the rapid publishing of explanations on inclusive announcements, movies, politics, and the economy. This growth has led to an increase in the breadth of topics covered. This emotion analysis includes many aspects. Arabic and OMCD survey big data were integrated with data from Twitter to inform this study. Different word embedding methods were implemented, such as Spacy (W2V), FastText, and Arabic Bidirectional Encoder Representation (AraBERT). In the context of sentiment analysis models, convolutional neural network (CNN), long short-term memory (LSTM), and recurrent neural network (RNN) were employed. The evaluation of model performance was based on accuracy. Deep learning (DL) methods using the AST (Arabic sentiment Twitter) dataset yielded 72% and 95% model accuracy rates. The accuracy rates for the OMCD (Offensive Moroccan Comments Dataset) is a dataset containing offensive comments in the Moroccan dialect. dataset fell within the range of 54% to 84%.

Keywords: Text Classification; Emoji Analysis; Sentiment Analysis; Deep Learning; Spacy; Arabic Language Review.

1. INTRODUCTION

In the contemporary era, social media offers an invaluable platform for expressing ideas and disseminating firsthand expertise about diverse events, commodities, and services. These valuable sources of information hold great appeal for online users in their pursuit of selecting the most optimal product or service for acquisition. Opinions possess significant merit due to their impartial and independent nature. Their foundation is based on actual customer experiences with a specific product or service [1].

User feedback is critical to businesses because it allows them to gauge consumer satisfaction and enhance their services. Considerable discussion has taken place about collecting data and discerning between positive and negative sentiments. The difficulties in learning Arabic have promoted more research on sentiment analysis of Arabic literature. The majority of sentiment analysis research is mainly concerned with the English language [2]. Recent significant and demanding improvements in Arabic Sentiment Analysis (ASA) have made it an essential subject of study [1]. A limited number of academic works address concepts, attitudes, feelings, and sentiments related to Arabic [2]. Sorting Arabic text according to its content into predefined categories is the primary goal of the ASA assignment. Notably, the way in which text is represented dramatically impacts how well AS can perform. Contextual embedding models facilitate obtaining universal sentence representations because they consider contextual factors and the meaning of words. Sentiment analysis, also known as opinion mining, is the process of determining whether an author has a favorable or unfavorable attitude about a particular topic [3].

However, identifying sentiment polarity is a difficult task that calls for sophisticated knowledge of natural language processing, information retrieval, linguistics, and a thorough comprehension of the textual context [4,5].

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Furthermore, analyzing and classifying these sentiments accurately is crucial because they can significantly impact many levels. Many studies in the existing literature use machine learning techniques to address sentiment analysis. Nevertheless, because deep learning (DL) has recently been more successful in various fields, data scientists and researchers have increasingly used DL to address related problems. Deep networks have proven to be efficient and effective in processing large datasets. Arabic is one of the four most utilized languages on the internet, with over 226 million users [6].

Billions of social media posts are generated daily, leading to an increased use of emoji and emoticons. Emoji and emoticons serve as invaluable resources for sentiment analysis. Sentiment analysis is a method employed to examine the text and discern various attitudes. Emoji have the potential to enhance sentiment analysis as well as other related endeavors, such as sarcasm detection. Its goal is to understand how emoji are used and the extent of their impact on detecting emotions [7]. The present study examined the sentiment analysis process in social media networks through the use of emoji. Twitter has been chosen as the primary information source for the analysis. Incorporating emoji characters increases the sentiment scores obtained through sentiment analysis. In addition, the impact of emoji characters is more noticeable in instances of positive opinions than negative ones [8]. The article concentrates on examining the sentimental roles of emoji in Arabic texts. Depending on the context, emoji can possess various sentimental roles. Emoji can function as enhancers, indicators, mitigators, reversers, or triggers. Emoji can also have a neutral impact on the sentiment of a text [9]. Sentiment analysis is the process of determining the attitudes expressed in a text; it has gained attention in various domains. This article discusses the difficulties of using emoji to categorize Arabic training data. It also presents an algorithm for handling tweets in Modern Standard Arabic and dialect Arabic. The authors highlight the complexity of the Arabic language and the need for more data to improve sentiment analysis [10]. The challenges to Arabic natural language processing are the variety of dialects and limited resources for hostile Arabic language detection. The challenges to Arabic natural language processing are the variety of dialects and limited resources for hostile Arabic language detection. This study addressed the problem of limited resources due to the use of the Arabic language, which has not been highlighted until very recently, as well as the large number of dialects and the limited use of classical Arabic. This problem is one of the most important problems that lead to differences in interpretations and conclusions when evaluating reviews using these methods. Models developed with MSA or a specific dialect often struggle when applied to other dialects due to linguistic differences [11]. The main contribution of this paper is that it provides a valuable analysis of the content of Arabic emotions with emoji and other dialects, with results comparable with those of previous studies.

2. Related Work

H. Cui, Y. Lin, and T. Utsuro [12] used emoji as training data to look at sentiment analysis of tweets. Certain emoji types appeared in tweets with inconsistent attitudes. Convolutional neural networks (CNNs) were used to classify sentiment. The CNN model performed better than the SVM model in sentiment analysis. The majority of the sentiment analysis studies of tweets obtained an equivalent number of positive and negative tweets using filters and manual judgment.

H. Abdellaoui and M. Zrigui [13] used the remote supervision method; emoji and sentiment lexicons from Twitter were used to gather and label a dataset for Arabic sentiment analysis. More than six million tweets were classified as neutral, harmful, or sound. The algorithm was capable of handling Modern Standard Arabic and dialect Arabic mixed-content tweets, but more training data were needed for better results with DL models like LSTM and CNN. An LSTM model trained on the head dataset also showed promising results for Arabic sentiment analysis. Furthermore, emoji were used as keywords for data gathering, and the issue of using dialect instead of Modern Standard Arabic was resolved. Finally, dialect words were replaced with synonyms in Modern Standard Arabic using a simple algorithm, and the top 20 most used emoji on Twitter, including the most subjective ones, were chosen.

Emoji express emotions on social media. Current sentiment analysis approaches using emoji have limitations. Therefore, this study by L. Yinxia et al. [14] proposed a bidirectional LSTSM sentiment analysis model. As such, the model incorporated the attention approach and achieved superior performance, and it showed that text had less of an impact on sentiment polarity than emoji. Moreover, compared with emoji-emb and SVM, the Bi-LSTM

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143
DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

(text emoji) model was better. Finally, the best model was the ea-bi-LSTM model, which effectively captured emoji influence; it achieved the highest f-scores in identifying sentiment polarities.

The entire process in the study done by Babu and Kanaga [15] used social media data for analysis and classification. It included textual data and emoji. The use of multi-class classification is more accurate than that of binary and ternary classification, which were employed in earlier studies. Intelligent techniques were used for apprehensiveness or dejection detection. Multi-class classification with DL algorithms showed a higher degree of precision. Combining different DL algorithms yielded a higher degree of precision. DL algorithms performed well on depression datasets. Combining CNN-LSTM algorithms gave a better precision of 92%. LSTM + FastText from Twitter, Kaggle scored 66%.

M. J. Althobaiti [16] suggested a technique for identifying hate speech and foul language in Arabic tweets. BERT, SVM, and logistic regression were contrasted for detection purposes. Furthermore, emoji and sentiment analysis were added features. The BERT-based model outperformed benchmark systems in every evaluation, and The addition of sentiment analysis enhanced the detection of hate speech and abusive language. The BERT model also achieved f1-scores of 84.3%, 81.8%, and 45.1% for offensive language recognition and hate speech detection. Although sentiment analysis improved the identification of abusive and hateful language, the influence of emoji as characteristics differed based on the size and uniqueness of the class.

P. Dey and S. Day [17] proposed a model for examining emotional tweet replies. The “tweepy” crawler data and machine learning techniques form the basis of the model. Based on the user's feelings, a sentiment is assigned to each tweet. The effect of emoticons on text mining is investigated. Adding emoticons raises the overall score for opinions. Sentiment analysis can benefit from an innovative viewpoint offered by emoji analysis.

L. Li and Wang [18] surveyed over 900 Weibo users to create a text-free emoji sentiment language. A total of 8,822 Weibo comments with indexed emoji were analyzed using the emotion lexicon. The findings revealed a public sentiment version of loss aversion, with negative emoji outperforming happy emoticons. Real-time observation and prediction of attitudes and reactions to social events can be achieved through emoji sentiment analysis. Emoji serve several purposes, as evident by the insights provided in [14] into several areas of study, including risk management, nonverbal communication, and public choice.

L. Berggren [19] examined two challenges: analyzing a non-English language and the impact of emoji. A sentiment-annotated dataset of Swedish texts with emoji was created, and a Swedish sentiment analysis model, called “Swevader,” was developed based on the English lexicon-based model “Vader.” The “Swevader” model achieved an accuracy of 53% and an f1-score of 47%. Emoji improved the analysis for most models, but classifying neutral texts took more work, raising concerns about using the dataset for model evaluation. Emoji contain valuable information and can be crucial when resources for text analysis are lacking. Therefore, emoji should be considered.

N. Nirbhik and S. Kumar [20] utilized opinion mining on Twitter to measure extremist affiliations, using emoticons to increase sentiment analysis accuracy. Transformers and BERT greatly influence NLP. Although the model performed well in predicting sentiment, room for improvement in identifying positive sentiment was found. A DL model was developed specifically for sentiment analysis on extremist-related tweets. The study showed that combining text and emoticons enhances classification accuracy. Future research could explore other language models and complex features. The potential for the model to be applied to other social media platforms was also considered.

3. Sentiment Analysis within Text and Emoji

Billions of social media posts lead to increased use of emoji and emoticons, which are valuable for sentiment analysis. Emoji can enhance accuracy and help detect sarcasm. The objective was to understand the use of emoji and its impact on sentiment analysis [21]. Using emoji characters in sentiment analysis on social networks,

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

particularly Twitter, resulted in higher sentiment scores, because emoji have a more substantial impact on positive opinions than negative opinions [22].

Emoji are popular tools for nonverbal communication in social media, providing insights into public sentiments. Positive emoji are used more frequently than negative ones, and negative emoji have more vigorous intensity. As a result, emoji indicate public sentiment in social communication [23]. A proposed model for sentiment analysis combines bidirectional LSTM and attention methods to document how emoticons affect the text. The model performed better than baseline models on a corpus of labeled “Sina Weibo” posts [24].

Emoji in Arabic texts can serve different sentiment roles depending on the context; they can emphasize, indicate, mitigate, reverse, or trigger sentiment. Emoji can also have a neutral effect on sentiment [25]. Emoji are also used in social media comments to convey emotions. As such, they were detected and analyzed using a dataset from Twitter, and accuracy rates for happiness, anger, and sadness were 80%, 72%, and 65%, respectively [26]. Sentiments conveyed by emoji differ between Arabic and European languages, particularly for emoji representing objects, nature, symbols, and specific activities. Food emoji also showed a high level of inconsistency regarding sentiments [27]. To improve sentiment analysis in the Arabic language, which is considered one of the most widely used languages on social networking sites, it was suggested that the study should be conducted on large data sets collected carefully and with caution [28].

4. Proposed Work

The primary function of the system is to analyze the sentiment (positive or negative) of Arabic-language comments. The practical parts of the system are explained in this section. The system has two essential components: gathering the dataset and using Arabic text for sentiment analysis; the system includes obtaining a substantial number of Arabic comment data as well. Compiling the dataset is a crucial stage. To guarantee the efficacy and precision of the system, the dataset must encompass a diverse range of themes and emotions. The second component of the system is a sentiment analysis of Arabic text; sentiment analysis aims to identify the feelings conveyed in a particular text. Here, the system prioritizes Arabic-language comments. As such, this section recommends using various technologies, such as DL algorithms and natural language processing techniques. The suggested system accomplished the ASA goal by utilizing DL methods. As shown in Figure (1), the procedure comprised reading the dataset, preparing it, and breaking the task into manageable chunks. The first section of the work used the Spacy method to encode emoji. In the second section, transformer models like AraBERT were used. The third section used various deep-learning methods, including CNN and LSTM. These models were all created using the two embedding techniques: FastText and Spacy.

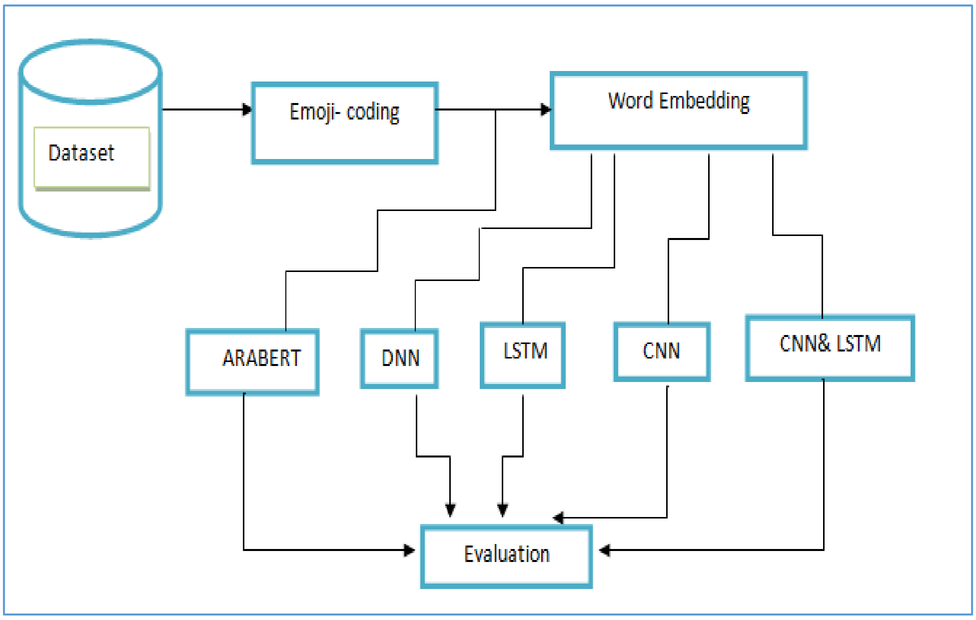

Figure (1): Block Schematic for the Suggested System

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

4.1 Dataset

The first dataset utilized in the investigation consisted of data derived from social media, and the information obtained from various social media platforms had undergone preliminary processing. The dataset encompassed textual data as well as emoji, and multiple datasets were employed, including tweets from diverse sources. This dataset was used for sentiment analysis, specifically focusing on an Arabic social media dataset encompassing 58,751 tweets from the Twitter platform. Additionally, 1,026 distinctive emoji were extracted from the Arabic dataset. The dataset comprised 29,849 positive tweets and 28,902 negative tweets. The dataset was partitioned with an 80:20 split ratio into a training set and a test set to make analysis easier. Ultimately, the dataset was employed to construct a classification model.

Classification methods are available that give better performance whenever the available data set is large, but this may lead to the emergence of the problem of imbalanced data. Many studies that address solutions to this problem have been presented [36].

The second dataset was in the Moroccan dialect. The reason for using data with limited resources such as Moroccan dialect data is to explore the findings and studies that have not been highlighted in these areas, and it also helps to distinguish the research from its counterparts as a result of presenting different aspects from others. which the researcher obtained from the GitHub website, contained 8,025 comments in this dialect, and it was also separated in the same proportions as the first data set. Two types of data were used according to the following table:

Table I Dataset Details

Dataset name	Website	Link	Num-of-Tweets	Arabic Dialects	lanced
Arabic-sentiment-twitter (AST)	Kaggle	https://www.kaggle.com/code/yasmeenhany/arabic-sentiment-analysis-using-arabic-bert - Arabic-sentiment-analysis [36]	58,751	Multi	Yes
OMCD	Github	https://github.com/kabilessefar/OMCD-Offensive-Moroccan-Comments-Dataset [11]	8,025	Single (Morocco only)	Yes

4.2 Emoji Coding and Preprocessing

The initial step in creating a successful program involves defining and quantifying emoji symbols. The researchers utilized digital coding methods, calculating the frequency, weight, and repetition of emoji in reviews. Through the Spacy methods, 284 emoji symbols were identified, while FastText was trained on extensive comments and texts to classify and study the impact of emoji on reviews. The examination of emoji in Arabic tweets revealed instances of inconsistent sentiment expression. To further analyze the relationship between emoji and sentiment, the researchers collected Arabic tweets with emoji and found that tweets with more emoji performed better. Then, the sentiment of tweets containing only one emoji was manually assessed, and consistent sentiment expressions in texts and emoji were observed. The frequency of symbols varied based on usage, resulting in four situations: opposite sentiment expression, emotionless tweets, ad-like tweets, and incomprehensible language.

The workflow algorithm analyzed emoji in a dataset by extracting them from each review and calculating their frequency. This information was then used to determine the weight of each emoji based on its occurrence in positive and negative reviews. A data frame was created to display the emoji along with their weights and classifications. In

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143
DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

addition, a vocabulary dictionary was designed for emoji and their weights. The pre-trained FastText model in Arabic was loaded and used to calculate the average contexts related to each emoji. Finally, the averages based on context were stored as expected quantities for each review. The other model (Spacy) was created using this same method, but the last method was more developed and trained in emoji and many Arabic dialects.

Figure (2): Represented emoji

Some studies have taken symbols as a basic element for evaluation, as they were studied as either a digital code or a code specific to each emoji, and most of them were converted into a text expressing the state, for example (☺) a smiling face or (☹) a sad face or other expressions, but we used a more accurate method because we encoded them (−1,1) to express them directly if they were negative or positive.

Emoji are very important because their presence in reviews greatly affects the results, as well as the possibility of relying on them to make a decision when not knowing the direction of the clear text in unclear and incomprehensible comments.

Preprocessing: The process of getting raw data ready for analysis or modeling is called dataset preparation. Data are cleaned, transformed, and persisted to improve usability and ease of use. Preprocessing is an important part of the data science journey because the quality of the data affects the efficiency of any subsequent analysis or modeling.

The preprocessing steps include:

• Removing all punctuation and diacritics from the text. These marks are distracting and might not have anything to do with how feelings are understood. Therefore, eliminating them before doing mood analysis is important.

• Making the Arabic text more consistent by changing some letters. Normalization means changing some letters in Arabic text to their normal form to get rid of mistakes and make the word embedding smaller. For instance, [رائئئئئئع] can be changed into [رائع].

• Removing any characters that appear more than once in the text. Characters that appear more than once can lead to mistakes when mood analysis is done.

Removing all Arabic words that are not negation characters: Words that contribute little or nothing to the meaning of the sentence can be removed without changing the overall tone of the text.

4.3 Word Embedding

4.3.1 FastText Embedding:

In 2016 [29], Facebook AI researchers unveiled FastText, an NLP library that demonstrated efficacy, and the primary goal of this endeavor was to augment the continuous skip-gram model [30] by integrating sub-word information. Their innovative approach involves acquiring word representations as the vector sum of the n-grams of the acquired character, due to the belief that sub-word information can be beneficial for rare words and the challenge of noisy data. Therefore, fastText word embedding was integrated into the experimental framework of the study as a possible text representation method. Within this framework, words were represented as sets of character n-grams, where n might have a value between 3 and 6, thereby capturing diverse aspects of words, ranging from suffixes to more extensive roots. This advancement, achieved through the fusion of DQN with deep learning, emphasized the importance of internal word information. For instance, when n equals three, the word (where) appears, which gives

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143
DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

the word the following expression: (wh, we, her, here, re, where). As is evident, the entire sequence is also provided. N-grams were found in the training corpus, and a vector representation was obtained for all of them. Finally, the word was expressed as the total of its n-gram vector representations.

4.3.2 Spacy Embedding:

Spacy [31] represents a neural network model and an open-source library for basic neural network activities. These tasks include tokenization, named entity recognition, similarity detection, sentence boundary identification, lemmatization, rule-based matching, tokenization, part-of-speech tagging, and other linguistic features. Also, the training phase of the stochastic model employs gradient and loss functions. Furthermore, the trained model performs well in scenarios involving transfer learning. Spacy is a freely available open-source software library that provides various practical tools for processing textual data. These tools include support for over 73 languages, state-of-the-art speed, a training system ready for production use, linguistically motivated tokenization, easy extensibility through custom components and attributes, and robust accuracy that has been rigorously evaluated.

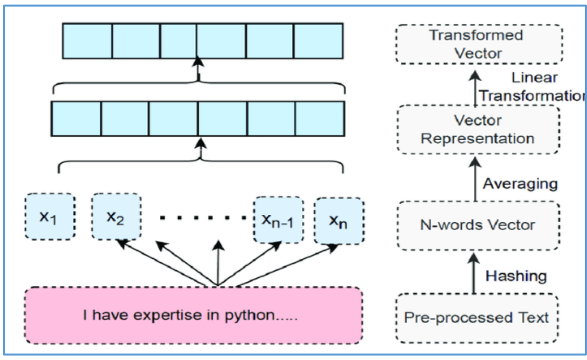

Figure (3) Architecture FastText Embedding [29]

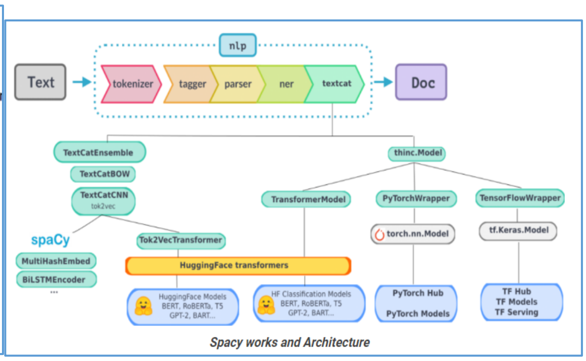

Figure (4) Spacy Architecture [31]

4.4 Deep Neural Networks (DNNs)

Data-driven modeling has made extensive use of DNNs. A DNN comprises layers with mathematical relationships between nodes and edges. Backpropagation is used to update these associations while the data are being trained. The modified relationships are then utilized as equations to forecast the output variables based on the training data. Consequently, a key benefit of DNNs is their capacity to convey the links present in the system despite its complexity and nonlinearity.

4.5 LSTM

Networks are particularly good at managing long-term dependencies and are made to remember data in longer sequences. In LSTM networks, the internal state of memory is driven by three gates: the input, forget, and output gates. These gates control which data in the memory should be added, removed, or left intact. The gates contain the amount of data that can flow through using a sigmoid layer and element-wise multiplication. The memory cell eventually learns the essential information based on the training process. LSTM attempts to solve some problems, including vanishing gradients, by imposing no bias against recent and evolving observations but keeping an error constant going back over a certain period [32].

4.6 CNN

Convolutional neural networks (CNNs) are feed-forward neural networks initially developed for computer vision applications [33, 34]. CNNs have shown notable progress in NLP tools. A layer within them exists that has locally applied convolutional filters. Unlike conventional neural networks, CNNs employ convolution instead of universal matrix multiplication. CNNs have become one of the DL algorithms with the fastest execution times due to reduced network complexity and weight count. Another advantage of CNNs is that they require less preprocessing. Their use

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

in several domains, including image processing, voice and handwriting recognition, and natural language processing, has been made possible by this characteristic.

4.7 AraBERT

A language model built on the BERT architecture is developed for Arabic processing. It is a crucial tool for Arabic natural language processing (NLP), enabling academics and industry experts to train models for various NLP tasks such as named entity recognition, text classification, and sentiment analysis. A vast collection of Arabic text, comprising a range of internet documents such as blogs, social media posts, and news stories, served as the initial training set for the AraBERT model. Similar to the pre-training process of BERT, the pre-training phase of AraBERT involved teaching it to infer words without including sentences. Using this paradigm, AraBERT could produce Arabic text representations of exceptional quality since it could comprehend the contextual links between words and their meanings [35]. AraBERT was tested in several NLP projects with promising results. For example, AraBERT surpassed previous pre-trained language models by attaining state-of-the-art performance in Arabic sentiment analysis [34, 35].

5. Experimental Results

After facing the difficulties of collecting data in Arabic and choosing the appropriate one, provided that it contains appropriate amounts of emoji to evaluate the studies, and after applying the various DL models and making changes to them by controlling the number of layers and choosing an appropriate design that shows the best results through experimentation and evaluation, these results were compared. Results from models with differences in including the word and experimenting with many methods and possibilities to reach the best outcomes were compared with the results reviewed in previous research; therefore, the aim was toward improvement and choosing the best-performing methods to obtain distinctive results in analyzing Arabic emotions and the dialects they contain. In this research, these results were revealed through many studies and experimenting with many methods and possibilities to find the best results compared to the results reviewed in previous research.

The following equation is used: Accuracy = (True Positives + True Negatives) / Total Samples.

The results were excellent, as two-word embedding methods were employed to compare DL models. As a result, the “Spacy” inclusion was better than “FastText” because it had been trained using a large number of languages and emoji, with approximately 1,026 emoji symbols learned, which yielded better results compared with “FastText.” In this work on DL models, the CNN model was found superior to the two methods in embedding. In addition, the transformer (AraBERT) was better than the other models, because it obtained 95%, which was the highest result compared to all other models. A previous study was conducted on the dataset, and BERT (transformer) was applied. The results were as described above.

Table II Accuracy Ratio for (AST) Arabic Sentiment Twitter Dataset

Accuracy (%)		**Model**
FastText	***Spacy-Word Vector***	
75%	90%	DNN
80%	92%	LSTM
83%	94%	CNN
74%	90%	CNN+LSTM (2L)
72%	87%	CNN+LSTM (4 Layers)
95%		AraBERT

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143
DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

90%	BERT [35]

The comparison presented in the preceding Table (4.8) demonstrates the current findings in relation to past research conducted on the identical dataset within the framework of a mixed model incorporating textual data and emoji. The findings indicate that the Spacy method exhibited superior performance by achieving the highest levels of accuracy in the classification of comments. Specifically, the analysis revealed that the highest accuracy rates were observed for the AraBERT and CNN models. Conversely, in the case of Fasttext, the LSTM model outperformed all others by achieving the highest accuracy rate. Juxtaposing these results with those of earlier studies showed that the accuracy rates of AraBERT and CNN surpassed those of the BERT model when applied to the same dataset.

Table III Accuracy Ratio for OMCD Dataset

Model	**Accuracy (%)**	
	Spacy-Word Vector	***FastText***
DNN	56%	55%
LSTM	54%	54%
CNN	56%	55%
CNN+LSTM(2L)	56%	54%
CNN+LSTM (4 Layers)	54%	54%
AraBERT+ Emoji coding	84%	
AraBERT [11]	81%	
LSTM [11]	78%	

In the OMCD dataset, the proposed models indicated a similar degree of effectiveness to the current best models, as evidenced in Table (4.14). (Antoun et al., 2020) projected an accuracy of 81% for their model, but our LSTM-based model predicted an accuracy of 78%, which was somewhat higher. This finding suggests that two possible methods for sentiment analysis in the OMCD dataset are the LSTM and AraBERT models.

The results obtained from the other dataset were relatively different from the previous study because the results were not lower. Nonetheless, AraBERT achieved the highest percentage by applying the two types of organizations mentioned previously. Consequently, the incorporation of fastText yielded a slightly improved examination of DL models, revealing that the LSTM+CNN model surpassed both approaches in embedding. The adapter (AraBERT) was also better than the other models because it obtained 84%, which was the highest result compared with all other models. A previous study was conducted on the data set, and AraBERT LSTM (transformer) was applied. The results were as described above.

6. Conclusion

This study examined the problem of the limited resources available to detect emotions in Arabic text and Twitter emoticons using embedding models. To address this issue, the OMCD dataset, which consists of feedback gathered from Moroccan YouTubers, was produced and made public. The researchers also conducted experiments utilizing several DL models and provided the findings to show the efficacy of our dataset. In addition, the significance of emoji was emphasized. In the OMCD dataset, AraBERT performed the best among all the other DL models. No previous studies have focused on analyzing Twitter to model the semantics of emoji within Arabic texts. The study compared word representation and embedding methods for sentiment analysis and DL models. The vectoral models of words and emoji were made available online. The models using the AraBERT transformer method resulted in the best performance. Moreover, the models using the “Spacy” word embedding method showed better performance

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

than FastText, and CNN showed the best performance among all DL models. Future studies plan to explore techniques and transformers to detect Arabic sentiment analysis, more sophisticated algorithms, sociolinguistic aspects, and ways to extract features more accurately. Regarding sentiment analysis, the CNN model outperformed either one of the models or both of them alone. Furthermore, the CNN-used model outperformed the CNN-non-static and CNN-static models by a small margin.

.References

[1]. B. Brahimi, M. Touahria and A. Tari, "Improving Sentiment Analysis in Arabic: A Combined Approach," *Journal of King Saud University-Computer and Information Sciences,* vol. 33, no. 10, pp. 1242-1250, 2021.

[2]. J. K. Alwan, A. J. Hussain, D. H. Abd, A. T. Sadiq, M. Khalaf and P. Liatsis, "Political Arabic Articles Orientation Using Rough Set Theory with Sentiment Lexicon," *IEEE Access*, vol. 9, pp. 24475-24484, 2021.

[3]. A. Khalid Al-Mashhadany, A. T. Sadiq, S. Mazin Ali and A. Ahmed, "Healthcare Assessment for Beauty Centers Using Hybrid Sentiment Analysis," *Indones. J. Electr. Eng. Comput. Sci.,* vol. 28, no. 2, p. 890, 2022.

[4]. A. A. Tuama and A. T. Sadiq, "Twitter Sentiment Analysis and Events Prediction Using Lexicon Sentiment Analysis for Iraqi Vernacular," *Solid State Technology,* vol. 63, no. 6, 2020.

[5]. E. Aydoğ and M. A. Akcayol, "A Comprehensive Survey for Sentiment Analysis Tasks Using Machine Learning Techniques," in *International Symposium on Innovations in Intelligent Systems and Applications (INISTA), IEEE*, pp. 1-7, 2016.

[6]. A. Omar, T. M. Mahmoud and T. Abd-El-Hafeez, "Comparative Performance of Machine Learning and Deep Learning Algorithms for Arabic Hate Speech Detection in OSNs," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), Springer International Publishing*, pp. 247-257, 2020.

[7]. B. Yoo and J. T. Rayz, "Understanding Emojis for Sentiment Analysis," in *The International FLAIRS Conference Proceedings,* vol. 34, Apr. 2021.

[8]. M. Shiha and S. Ayvaz, "The Effects of Emoji in Sentiment Analysis," *Int. J. Comput. Electr. Eng. (IJCEE),* vol. 9, no. 1, pp. 360-369, 2017.

[9]. Y. Lou, Y. Zhang, F. Li, T. Qian and D. Ji, "Emoji-Based Sentiment Analysis Using Attention Networks," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP),* vol. 19, no. 5, pp. 1-13, 2020.

[10]. M. J. Althobaiti, "BERT-Based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis," *International Journal of Advanced Computer Science and Applications,* vol. 13, no. 5, 2022.

[11]. K. Essefar, H. Ait Baha, A. El Mahdaouy, A. El Mekki and I. Berrada, "OMCD: Offensive Moroccan Comments Dataset," *Language Resources and Evaluation,* vol. 57, no. 4, pp. 1745-1765, 2023.

[12]. H. Cui, Y. Lin, and T. Utsuro, "Sentiment Analysis of Tweets by CNN Utilizing Tweets with Emoji as Training Data," in *Proc. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM),* pp. 1-8, 2018.

[13]. H. Abdellaoui and M. Zrigui, "Using Tweets and Emojis to Build TEAD: An Arabic Dataset for Sentiment Analysis," *Computación y Sistemas,* vol. 22, no. 3, pp. 777-786, 2018.

[14]. P. Wang, F. Zhou, K. Guan and W. Tan, "In Vivo Therapeutic Response Monitoring by A Self-Reporting Upconverting Covalent Organic Framework Nanoplatform," *Chemical Science,* vol. 11, no. 5, pp. 1299-1306, 2020.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[15]. N. V. Babu and E. G. M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review," in *SN Computer Science, Springer,* vol. 3, no. 1, 2022. doi:10.1007/s42979-021-00958-1.

[16]. M. J. Althobaiti, "BERT-Based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis," *International Journal of Advanced Computer Science and Applications,* vol. 13, no. 5, 2022.

[17]. P. Dey and S. Dey, "Sentiment Analysis of Text and Emoji Data for Twitter Network," *Al-Bahir Journal for Engineering and Pure Sciences,* vol. 3, no. 1, 2023. doi:10.55810/2313-0083.1034.

[18]. I. Li and X. T. Wang, "Nonverbal Communication with Emojis in Social Media: Dissociating Hedonic Intensity from Frequency," *Language Resources and Evaluation,* vol. 57, no. 1, pp. 323–342, 2023. doi:10.1007/s10579-022-09611-6.

[19]. L. Berggren, "Sentiment Analysis for Swedish: The Impact of Emojis on Sentiment Analysis of Swedish Informal Texts," M.Sc. Thesis, *Umea University*, Umea: Sweden, 2023.

[20]. N. Nirbhik and S. Kumar, "Integrating Text and Emoticons for Detecting Extremist Affiliations on Twitter Using Deep Learning," in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), IEEE,* pp. 1-6, Oct. 2023.

[21]. S. Ayvaz and M. O. Shiha, "The Effects of Emoji in Sentiment Analysis," *International Journal of Computer and Electrical Engineering,* vol. 9, no. 1, p. 360, 2017. doi:10.17706/ijcee.2017.9.1.360-369.

[22]. I. Li and X. T. Wang, "Nonverbal Communication with Emojis in Social Media: Dissociating Hedonic Intensity from Frequency," *Language Resources and Evaluation,* vol. 57, no. 1, pp. 323–342, 2023. doi:10.1007/s10579-022-09611-6.

[23]. I. O. U. Yinxia and J. Zhang, "Emoji-Based Sentiment Analysis Using Attention Networks," *ACM Transactions on Asian and Low-Resource Language Information Processing,* vol. 19, no. 5, 2020. doi:10.1145/3389035.

[24]. S. A. A. Hakami, R. J. Hendley and P. Smith, "Emoji Sentiment Roles for Sentiment Analysis: A Case Study in Arabic Texts," in *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*,* pp. 346-355, Dec. 2022.

[25]. A. B. Nassif, A. Elnagar, I. Shahin and S. Henno, "Deep Learning for Arabic Subjective Sentiment Analysis: Challenges and Research Opportunities," in *Applied Soft Computing, Elsevier Ltd,* vol. 98, 2021. doi:10.1016/j.asoc.2020.106836.

[26]. P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Sub Word Information," *Arxiv Preprint*, 2016. arxiv:1607.04606.

[27]. T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Arxiv Preprint*, 2013. arxiv:1301.3781.

[28]. A. A. R. Hussien, and N. A. Z. Abdullah, "A Review for Arabic Sentiment Analysis Using Deep Learning," *Iraqi Journal of Science, University of Baghdad-College of Science,* vol. 64, no. 12, pp. 6572-6585. doi:10.24996/ijs.2023.64.12.37.

[29]. C. Chantrapornchai and A. Tunsakul, "Information Extraction Tasks Based on BERT and SpaCy on Tourism Domain," *ECTI Transactions on Computer and Information Technology (ECTI-CIT),* vol. 15, no. 1, pp. 108-122, 2021.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 132-143

DOI: https://doi.org/10.25195/ijci.v50i2.534
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[30]. A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Short Papers,* 2017.

[31]. 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," *Social Network Analysis and Mining,* vol. 10, no. 1, 2020.

[32]. W. Antoun, F. Baly and H. Hajj, "AraBERT: Transformer-Based Model for Arabic Language Understanding," *Arxiv Preprint*, 2020. arxiv:2003.00104.

[33]. H. S. Abdullah, N. H. Ali and N. A. Z. Abdullah, "Evaluating the Performance and Behavior of CNN, LSTM, and GRU for Classification and Prediction Tasks," *Iraqi Journal of Science,* vol. 65, no. 3, pp. 1741-1751, 2024. doi:10.24996/ijs.2024.65.3.43.

[34]. R. Bensoltane and T. Zaki, "Towards Arabic Aspect-Based Sentiment Analysis: A Transfer Learning-Based Approach," *Social Network Analysis and Mining,* vol. 12,pp. 1-16, 2022.

[35]. S. A. Razoqi and G. A. A. Al-Talib, "A Survey Study on Proposed Solutions for Imbalanced Big Data," *Iraqi Journal of Science,* vol. 65, no. 3, pp. 1648-1662, 2024. doi:10.24996/ijs.2024.65.3.37.

[36]. https://www.kaggle.com/code/yasmeenhany/arabic-sentiment-analysis-using-arabic-bert-Arabic-sentiment-analysis.