Journal of Computing and Communication, Vol. 3, No.2, PP. 10-22, 2024

Arabic Sentiment Analysis using Deep Learning and Machine Learning approaches.

Abdel Nasser H. Zaied [a], Gawaher Soliman [b]

[a] Faculty of Computer Science, Misr International University, Cairo, Egypt

[b] Faculty of Computers and Informatics, Zagazig University, Al-Sharqia, Egypt

*Corresponding Author: Gawaher Soliman [***gawahersoliman@zu.edu.eg***]

ARTICLE DATA

Article history:
Received 5 Feb 2024
Revised 1 June 2024
Accepted 24 June 2024
Available online

Keywords:
Sentiment Analysis
Sentiment Classification
Convolutional Neural Networks
Long-short Term Memory networks, Logistic Regression
Random Forest

ABSTRACT

Sentiment analysis is defined as an analysis of text to determine the sentiment expressed within it. This text emphasizes the significance of sentiment analysis in web mining and data classification, with detailed illustrations on sentiment analysis of the Arabic language. This study proposed a sentiment analysis framework to review the Arabic text. Two textual representations were explored: term frequency-inverse document frequency (TF-IDF) and word embedding via Word2vec. Various methods have been suggested for categorizing sentiments in Arabic text based on a dependable dataset, including Long Short-Term Memory (LSTM), hybrid LSTM-CNN, Convolutional Neural Network (CNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), and Random Forest (RF). The findings indicated that these methods enhanced Accuracy, precision, Recall, and F1-score. The LR and SVM classifiers accomplished the highest Accuracy with 87%, while the other classifiers (LSTM), (CNN-LSTM), (CNN), (MNB), (RF), and (DT) achieved accuracies with 86.41%, 86.10%, 85.26%, 85%, 84% and 81% respectively.

1. Introduction

Social media platforms, including Facebook and Twitter, have become significant conduits for obtaining public sentiment data about specific products or subjects. Recently, Arabic has gained recognition as one of the most widely disseminated languages on social media platforms, particularly Twitter and Facebook.

Businesses and governments utilize social media to gather and analyze user attitudes, including in different languages like Arabic. Conducting sentiment analysis in Arabic requires various preprocessing tasks, including morphology, phonetics, sentence segmentation, semantic analysis, part-of-speech tagging, named entity recognition, symbolic analysis, subjective analysis, and manual annotation utilizing lexicons or a corpus. [1].

Sentiment analysis (SA) involves natural language processing (NLP), the computational study of human languages from a scientific perspective. It is a method of capturing the feelings or opinions of people towards a specific topic. The literature has identified three main SA approaches, namely the hybrid approach, the lexicon-based strategy, and machine learning, for Arabic and English languages [2].

Arabic is among the most widely spoken languages in the globe. It was classified primarily into three categories: Modern Standard Arabic (MSA) serves as the authoritative medium of communication in a formal setting. Traditional or classical Arabic, like old Arabic manuscripts and religious Arabic scripts such as the holy Quran language. Dialectical Arabic (DA) (Colloquial Arabic or Slang Arabic), is commonly spoken among people every day [3]. Analyzing sentiment in Arabic involves extracting and comprehending the emotions or feelings in Arabic text.

Limited research has been devoted to the analysis of sentiment in Arabic. This is due to the scarcity of annotated corpora for the Arabic language and the challenging processing of the Arabic natural language [2, 4, 5, & 6].

This paper introduces new approaches to Arabic sentiment analysis that utilize (CNN), (LSTM), (Hybrid CNN-LSTM) (LR), (SVM), (MNB), (RF), and (DT). Also, it discussed the importance of sentiment analysis in (NLP) and the need for various approaches to achieve accurate classification. The suggested sentiment analysis strategy involves the following steps: preprocessing the text, collecting data, selecting the model, extracting features, training the model, and evaluating and analyzing the model. The paper's significance lies in its contribution to Arabic Sentiment Analysis. by proposing new approaches and highlighting the value of accurate sentiment analysis in various applications. Also, the findings of this study can be valuable for researchers working in Arabic sentiment analysis.

The following sections of the paper are organized in the following manner: Section 2 provides a comprehensive review of previous studies conducted in Arabic Sentiment Analysis. In contrast, Section 3 presents a detailed explanation of the proposed work. Section 4 provides an account of the experimental research, including the findings and subsequent analysis. Finally, Section 5 includes the conclusion and recommendations for future research.

2. Works of Relevance

Platforms like Twitter and Facebook are commonly used as data sources for sentiment analysis, which involves analyzing users' opinions, interests, and behaviors related to global events [2 & 7]. In Arabic sentiment analysis, deep learning and machine learning approaches have been successfully applied in several studies to determine Arabic sentiment. as shown in Table (1).

2.1 Using advanced deep learning techniques to analyze sentiments in Arabic, Al-Smadi et al. [8], developed a model for analyzing sentiment in Arabic hotel reviews using Aspect-Based Sentiment Analysis (ABSA). They extracted various linguistic features from 2,291 reviews using AraNLP and MADAMIRA tools. Then, they employed a Recurrent Neural Network (RNN) model implemented with the Deep Learning 4j2 Framework, and SVM was implemented with The Weka 2.7 java-based model. The SVM accomplished an accuracy of 95.4%, and the RNN accomplished 87% accuracy in sentiment classification in the proposed model.

Ammar and Rania [9] proposed a deep learning (DL) method for Arabic sentiment analysis utilizing CNN, LSTM, and RCNN models. Their findings from the experiment indicated that LSTM attained an average accuracy of 81.31%, surpassing the performance of RCNN and CNN. Additionally, applying data augmentation to the dataset resulted in an 8.3% increase in LSTM accuracy, as they discovered. Al-Smadi et al. [10] also suggested a deep learning model utilizing the LSTM network to analyze ABSA (Aspect-Based Sentiment Analysis) in Arabic. The model predicts the sentiment expressed in an aspect by calculating the phrase's attention weight feature vector and desired aspect. This model has accomplished an accuracy of 82.6% on the Arabic Hotels reviews dataset.

Abdulhakeem et al. [11], suggested a model that tackles Arabic Sentiment Analysis (ASA) by using Long Short-Term Memory (LSTM). The results show an accuracy of about 82% and an F-score of about 81.6%.

In their study, Alharbi et al. [12], developed a deep-learning model that used GRU, LSTM, and ensemble methods to evaluate Arabic sentiment. They assessed the model's effectiveness by utilizing six datasets and attained a 94.32% accuracy, surpassing the other art models.

In addition, Asma and Zouhour [13] have created a model that combines a lexicon-based approach with a customized sentiment rule-based engine called VADER. When tested with the Tunisian Arabic Dialect, the hybrid model had a commendable performance, with an accuracy of 85% in classification. Ahmed [14] proposed a Sentiment Analysis (SA) approach using a bi-LSTM model. The findings indicate that the developed method has an accuracy of 76.05% with a Macro-F1 score.

Elhassan et al. [4] utilized Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and a combination of CNN and LSTM (hybrid CNN-LSTM) to do sentiment analysis on Arabic text. These models utilized the derived characteristics from word embeddings like fastText and Word2Vec. The models were evaluated using two well-established Arabic datasets: the Hotel Arabic Reviews Dataset (HARD) for

hotel reviews and the Large-Scale Arabic Book Reviews (LARB) for book reviews, spanning several setups. The results indicated that the CNN model attained the best Accuracy of 94.69%, closely accompanied by the CNN-LSTM model at 94.54% and the LSTM model at 94.63%.

2.2 Using Machine learning techniques to analyze sentiments in Arabic

In a study by Al-Smadi et al. [15], an improved technique for Arabic hotel reviews sentiment analysis (ABSA) was proposed. The researchers employed monitored machine learning techniques involving SVM, DT, NB, KNN, and Bayes Networks. The assessment results indicated that all of the classifiers in their methodology performed better than the standard method, with the SVM algorithm accomplishing the utmost Accuracy of 95.4% compared with the rest of the algorithms.

Elshakankery and Ahmed [16] developed a hybrid Arabic sentiment analysis (ASA) system that combines lexicon-based and machine-learning approaches. Experiments were carried out using various datasets such as ASTD and ArTwitter, and SVM, LR, and RNN classifiers were utilized. The LR classifier demonstrated the most outstanding performance with an accuracy of 83.73%, outperforming the RNN at 81.62% and the SVM at 81.52%. In this text, Gamal et al. [17] conducted experiments utilizing various machine learning algorithms, such as NB, SVM, Ridge Regression (RR), and AdaBoost, on a dataset comprising 151,548 tweets. They utilized TF-IDF for feature extraction and found that Ridge Regression (RR) demonstrated outstanding performance with an accuracy rate of 99.9%.

Alyami and Olatunji [18] studied sentiment classification in Arabic using the SVM model. Their primary objective was to classify sentiments into negative or positive categories. The investigation used a Twitter dataset covering several societal topics in Saudi Arabia. The results demonstrated the significant efficacy of the Support Vector Machine (SVM) algorithm, with an accuracy rate of 89.83%. Alsalman, a 19-year-old, created a Discriminative Multinomial Naïve Bayes (DMNB) model to categorize Arabic tweets as either positive or negative. The model was assessed using a 5-fold cross-validation approach and compared against other machine learning methods using the identical dataset. The results suggested that the DMNB model had an impressive accuracy of 87.2%. Alharbi and Qamar [20] conducted a study on sentiment categorization of Arabic reviews about coffee shops and eateries. They compared multiple machine learning algorithms and determined that the SVM method exhibited % the highest precision of 89.0% among the tested methods.

Govindan and Balakrishnan [21], utilized four machine learning techniques: SVM, NB, DT, and RF, to analyze a dataset that was manually converted into a numerical format. The methods employed aimed to identify sarcasm in negative tweets using supervised machine-learning strategies. The proposed model achieved an average accuracy of 75%. In addition, Musleh et al. [22] conducted a study on Arabic sentiment analysis in the context of depression. The researchers utilized six machine learning techniques and determined that the Random Forest (RF) classification algorithm had the highest level of Accuracy, with a score of 82.39%, surpassing the performance of other algorithms. Maria and Bdulla [23] conducted a study to assess the influence of two feature selection methods, Chi-Square and Information Gain (IG), on the effectiveness of three classifiers (DT, KNN, and SVM) trained on Jordanian Arabic Tweets to develop a sentiment analysis model. The study found that the Information Gain Algorithm outperformed the Chi-Square Algorithm in the feature selection. The outcome was a decrease of 61% in the quantity of features and an improvement of 10% in the Accuracy of the classifier. In all studies, the SVM classifier outperforms the DT and KNN models, obtaining the most excellent Accuracy of 85% using the IG method.

Ultimately, Musleh et al. [5] created a Natural Language Processing (NLP) model to categorize Arabic comments as favorable or unfavorable. The model employs six classifiers: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, K-Nearest Neighbours (KNN), Decision Tree, and Random Forest. The model underwent training using a dataset consisting of 4212 comments that were labeled. The training process resulted in a Kappa score of 0.818. The NB algorithm yielded an accuracy of 94.62% and an MCC score of 91.46%. The Naive Bayes algorithm's precision, Recall, and F1-measure values were 94.64%,

94.64%, and 94.62%, respectively. The Decision Tree exhibited unsatisfactory performance, achieving an accuracy of 84.10% and an MCC score of 69.64% in the absence of TF-IDF. This research offers valuable insights for content providers seeking to improve their material and engage with their audience by examining viewers' emotional responses to videos.

TABLE 1: An overview of the studies surveyed by Arabic-level feature.

Reference	Classification Algorithm	Results
Deep Learning Approaches		
8 (2018)	RNN	accuracy of 87%.
9 (2019)	CNN, LSTM, and RCNN	accuracy of 81.31%.
10 (2019)	LSTM	accuracy of 82.6%.
11 (2020)	LSTM	accuracy of 82%
12 (2021)	LSTM	accuracy of 94.32%.
13 (2023)	hybrid model	accuracy of 85%.
14 (2023)	bi-LSTM	accuracy of 76.05%
4 (2023)	CNN, LSTM, and CNN-LSTM	CNN achieved the highest Accuracy of 94.69%, followed by LSTM (94.63%), and CNN-LSTM (94.54%)
Machine Learning Approaches		
15 (2018)	KNN, SVM, NB, DT, and BN	SVM achieved the highest Accuracy of 95.4%.
16 (2019)	SVM and LR,	LR with an accuracy of 83.73%, followed by SVM with 81.52%.
17 (2019)	SVM, NB, RR, and AdaBoost	accuracy of 99.9%.
18 (2020)	SVM	accuracy of 89.83%.
19 (2020)	DMNB	accuracy of 87.2%.
20 (2021)	SVM	accuracy of 89.0%.
21 (2022)	SVM, NB, DT, and RF	accuracy of 75%.
22 (2022)	RF	accuracy of 82.39%.
23 (2022)	DT, KNN, and SVM	accuracy of 85%
5 (2023)	SVM, NB, LR, KNN, DT, and RF	NB achieved the highest Accuracy of 94.62% and an MCC score of 91.46%, DT had a suboptimal performance with 84.10%

3 Research methods

The following section addresses the collection of the dataset, preprocessing, extraction of features, and generation of the model. The proposed methodology for building a sentiment analysis approach is shown in Figure 1.

3-1 Data Collection

Regarding the methodology utilized to analyze sentiment, data collection is the first step involving various methods for collecting textual data, such as extracting tweets or reviews using Application Programming Interfaces (APIs) and using existing datasets from websites such as Kaggle. Accurate data collection is essential as subsequent steps in the sentiment analysis process depend on it. The Arabic Companies Reviews For Sentiment Analysis dataset from Kaggle [24] was used in this research. The dataset includes reviews of eight Arabic companies: Talbat, Hotels, Alahli_bank, Swvl, Telecom_Egypt, Raya, Venus, and Hilton.

The dataset contains 67127 reviews, divided into 20339 positive reviews, 23753 neutral reviews, and 23035 negative reviews. After removing the duplicated reviews, the dataset becomes 65236 reviews

which are divided into 23462 neutral reviews, 22707 negative reviews, and 19067 positive reviews, as shown in figure (2). The reviews of the eight Arabic Companies are shown in Figure (3) and Table (2).

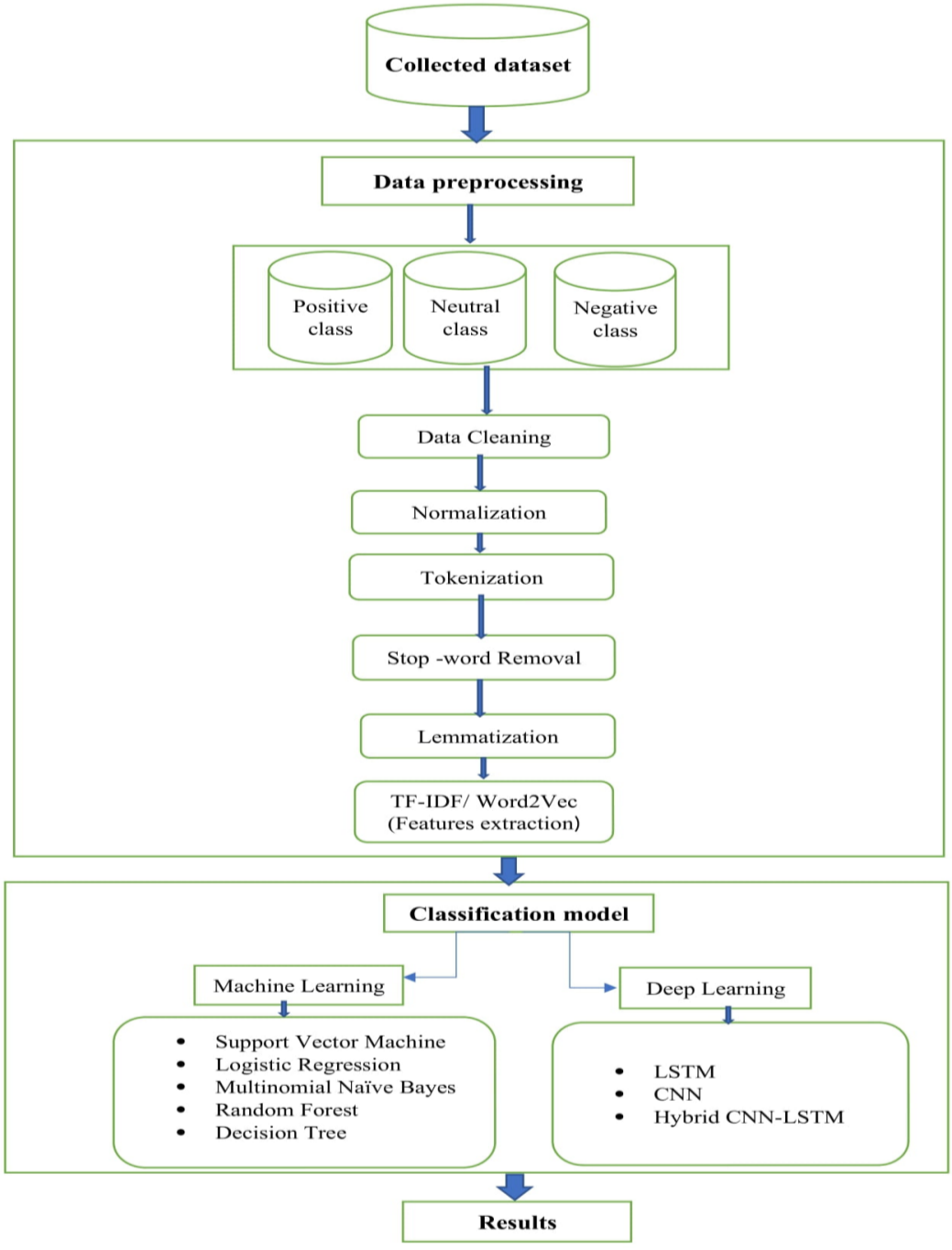

FIGURE 1. Structure workflow of the proposed work

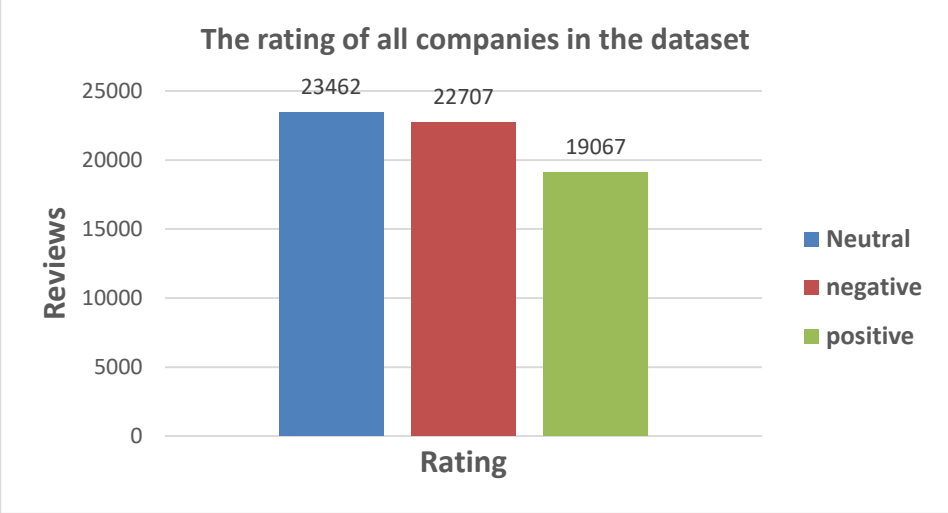

FIGURE 2. The rating of all companies in the dataset after removing duplication

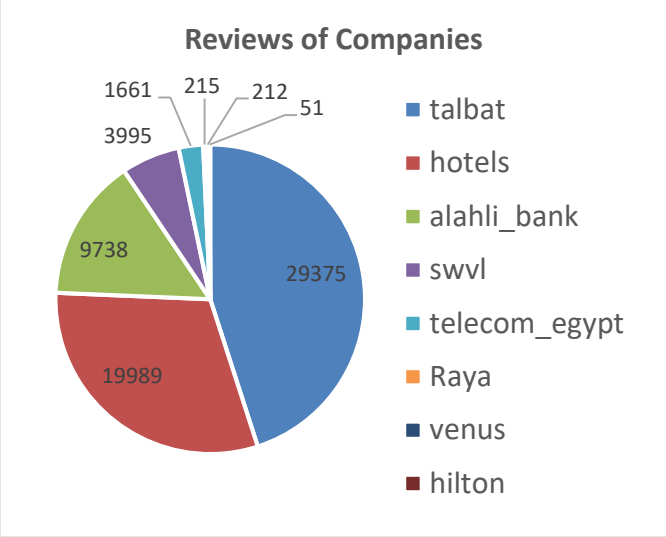

FIGURE 3. The reviews of the eight companies in the dataset

TABLE 2: The reviews of the companies in the dataset

Company	Number of Reviews
Talbat	29375
Hotels	19989
Alahli_bank	9738
Swvl	3995
Telecom_Egypt	1661
Raya	215
Venus	212
Hilton	51

3-2 Text Preprocessing

Text preprocessing plays a vital role in analyzing the sentiment, as it helps clean and normalize the text data, making it easier to interpret. The preprocessing step involves several techniques: data cleaning, tokenization, stop word removal, normalization, and stemming or lemmatization.

- **Data Cleaning:** Data Cleaning is the preliminary step to eliminate special characters, numbers, exclamation, and punctuation marks such as **('!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~')**, URLs, diacritics, mentions superfluous spaces and many other symbols, foreign words, Removing duplicated data, removing non-Arabic rows, words, letters and digits, and discarding any other rows that are written in the English language or other languages as only focusing on all Arabic tweets that were belonging to specific terms from the comment.

يقول ادخل رقم الهوية ادخلها يقول اسف انهاء؟؟؟	original review
يقول ادخل رقم الهوية ادخلها يقول اسف انهاء	cleaned review

- **Tokenization:** Tokenization refers to segmenting a body of text into smaller units known as tokens, which typically consist of individual words or short sentences.

- **Stop Word Removal**: Stop words are frequently used words in a language that don't contribute to t he overall meaning of a sentence, such as "و" (and), "في" (in), and "على" (on).

جميل ومريح وسهل التعامل	original review
جميل مريح سهل التعامل	after stop word removal

- **Normalization**: some Arabic letters can be normalized and exchanged into another form and then written directly in many ways, such as: ('ئ'-'ى' => 'ي') ('ة'-'ه' => 'ه') ('آ'-'إ'-'أ' => 'ا')

بعد التحديث الأخير أصبح	original review
التحديث الأخير	after normalization

- L**emmatization:** Considering the context, the process involves determining the meaningful base form of a word, referred to as its lemma. It is essential to acknowledge that a single word may possess multiple lemmas. We must also recognize the part of speech (POS) tag linked to the word in that context. Stemming gets more idol performance than lemmatizing, but the second gets higher Accuracy than the first. Lemmatizing takes more time, effort, and resources than stemming.

ممتاز الطلب وسرعة التوصيل	original review
ممتاز طلب سرع توصيل	after lemmatization

A sample of data preprocessing step is showing in table (3).

TABLE 3: A sample of data preprocessing step

Stages	Sample Review
The Original tweet	أسوأ تجربة &أسوأ خدمة على الإطلاق
Cleaning	أسوأ تجربة أسوأ خدمة على الإطلاق
Normalization	اسوا خدمة على الاطلاق اسوا تجربة
Removing Stop-words	اسوا خدمة الاطلاق اسوا تجربة
Lemmatization	اسوا خدم الاطلاق اسوا تجربة

3.3. Feature Extraction

This study seeks to analyze the sentiment on Arabic Companies' Reviews by utilizing deep learning and machine learning methodologies. We utilized TF-IDF and word embedding (word2vect) to identify the most impactful characteristics of the model and attain the utmost Accuracy possible.

We utilized five distinct monitored machine learning models, namely SVM, RF, LR, DT, and MNB, alongside three deep learning models., specifically LSTM, CNN, and hybrid CNN-LSTM.

3.3.1. Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF approach is used for text organization and extracting features. The term frequency refers to the frequency with which a word is used in a particular document. In contrast, the inverse document frequency counts the frequency of a word's appearance over the entire document, showing its level of commonality or rarity [5, 6].

The TF-IDF is part of the Scikit-learn library, and it is calculated using the following equation:

$$Wij = tf_{ij} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

The variable tf_{ij} indicates the frequency of word i in document j, whereas dfi represents the number of documents that contain word i. N indicates the total number of documents.

3.3.2. Word embedding

The distributional hypothesis is the basis of Word embeddings. It suggests that words exhibiting similar contexts are likely to have analogous meanings. Word embeddings are fixed-length distributed vector representations of words, which capture their textual meanings through real-value vector mappings. According to [25], similar words are associated with similar vectors.

Word embeddings have exhibited enhanced performance across various Natural Language Processing (NLP) tasks when employed as the foundational input representation. They have proven their effectiveness in classifying raw text, such as English [26&27], Chinese [28], Arabic [29], and more. The generation of this mapping involves using different methods, such as neural networks, reducing dimensions on word co-occurrence matrices, and using probabilistic models. Word2vec, a widely utilized and efficient prediction-based technique for word embedding, was employed [30]. It is known for its effectiveness in terms of both time and space. Word2vec consists of a neural network with two layers. Within this configuration, the document serves as the input, while the output comprises a collection of feature vectors with absolute numerical values. Each vector corresponds to a word and possesses a distinct, unchanging dimension. The language modeling problems utilize two main learning algorithms: Skip-Gram (SG) and Continuous Bag-of-Words (CBOW). The skip-gram approach requires predicting the words that appear in the context of a central word, which leads to more computing complexity. On the other hand, CBOW involves predicting the current word by considering a window of surrounding words.

In this study, we trained the model in Python using Gensim package. We used the skip-gram with a vector dimension of 100, the minimum word count is 10, epochs are 100, and sg is 1. Also, we used padding to ensure that sequences of variable lengths have consistent dimensions by adding special values or tokens to the sequences so that they all have the same length. For padding we used the following parameters (padding='post', maxlen=20, truncating="post", dtype='float32')

4. Evaluation metric and results

This section addresses environment setup, used parameters, evaluation metrics, and experimental findings.

4.1 Environment Setup and used parameters.

In this study, the most suitable environment setup for a proposed framework is Python, which is used to evaluate the performance due to its extensive ecosystem of application programming interfaces (APIs). We used a Lenovo Legion 5 laptop for the experimental setup with the following features: Nvidia Geforce(RTX 3060) graphics card with 6GB of RAM, 12 Generation Intel (R) processor Core(TM)i7 -12700H (20) CPUs, 16 GB RAM, Hard disk of 1TB SSD and windows 10 pro-64 bit. We used Anaconda for Python.

The experimentation involved training LSTM, CNN, hybrid CNN-LSTM, LR, SVM, MNB, RF, and DT models using a training set and testing with a wide range of hyperparameters and settings to find the best parameters. The number of epochs is a hyperparameter that determines how many complete passes the classifier makes through the training dataset; all deep learning models trained 50 epochs. A dropout rate of 0.1 is employed to prevent overfitting, which helps regularize neural networks. A filter size of 3 is used for the CNN model, accompanied by a ReLU activation function layer. We utilized a softmax layer containing three output units to ascertain the sentiment reviews' polarity, including positive, neutral, and negative categories. We used the Adam Optimizer with sparse categorical cross-entropy optimization in all experimentations. Table (4) shows the optimized DL model experiment hyperparameters.

TABLE 4: The Hyperparameters of DL Models

DL Models	Dropout	Activation	Activation Output
CNN	0.1	ReLU	Softmax
LSTM	0.1	ReLU	Softmax
CNN-LSTM	0.1	ReLU	Softmax

4.2. Evaluation Metrics

Evaluation Metrics were aimed at declaring relationships in sentiment analysis, which can be represented in Accuracy, Precision, Recall, and F1-score as the following:

- Accuracy: It presents how frequently the correct model predicts sentiment or opinion ratings [28]. It can be calculated by:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

As TN, TP, FN & FP represent true negative, true positive, false negative & false positive consecutively.

It is necessary to calculate both precision and recall values in data mining. They are used in both retrieved and relevant document conditions [30]. It can be despised as the following:

- $Precision = \frac{TP}{TP+FP}$ (3)
- $Recall = \frac{TP}{TP+FN}$ (4)

- F1-score: It is one of the statistical analyses for the binary classification. It is applied to measuring the trial accuracy [30]. It can be calculated by:

$$F_1 \text{ score} = \frac{2}{\frac{1}{recall}+\frac{1}{precision}} \quad (5)$$

4.3. Experimental Results

The training set comprised 80% of the dataset, while the testing set comprised 20%. Using a training set, the experiments start training the dataset with the applied DL models of LSTM, CNN, and hybrid CNN-LSTM. We used word embedding (Word 2Vec) to deal with this task. Word embedding converts the training set into a vector of integers (vectors). We used the Gensim package to train the model in Python. We also applied five ML models, LR, RF, SVM, DT, and MNB, to evaluate that framework. We used frequency * Inverse Document Frequency (TF*IDF) with the five machine-learning classifiers. Finally, testing the model classifies the sentiment polarity of the dataset automatically by monitoring models on how to identify sentiments into positive, negative, and neutral reviews. We computed Precision, Recall, F1-score, and accuracy measures. The outcomes of the suggested framework are displayed in Tables (3) and (4), illustrating the performance of DL and ML models, respectively.

TABLE 5: The Results of DL Models

DL Models	Accuracy (%)
CNN	85.26
LSTM	86.41
CNN-LSTM	86.10

TABLE 6: The Results of ML Models

ML Models	Accuracy (%)	Precision	Recall	F1-score
LR	87	87.33	87	87
SVM	87	87	87	86.66
MNB	85	84.66	84.33	84.66
RF	84	84.33	83.66	83.66
DT	81	81	80.66	80.66

The LR and SVM models both attained the highest Accuracy at 87%, while the remaining models achieved accuracies of 86.41%, 86.10%, 85.26%, 85%, 84%, and 81% using (LSTM), (CNN-LSTM), (CNN), (MNB), (RF), and (DT) respectively. Figures 4 to 8 depict the confusion matrix of the machine learning models, while Figure 9 illustrates the Accuracy of the models used. The results suggest that the SVM and LR models perform better than the deep learning models. However, when trained with larger datasets, the deep learning models exhibit higher performance, indicating a positive relationship between dataset size and accuracy improvement.

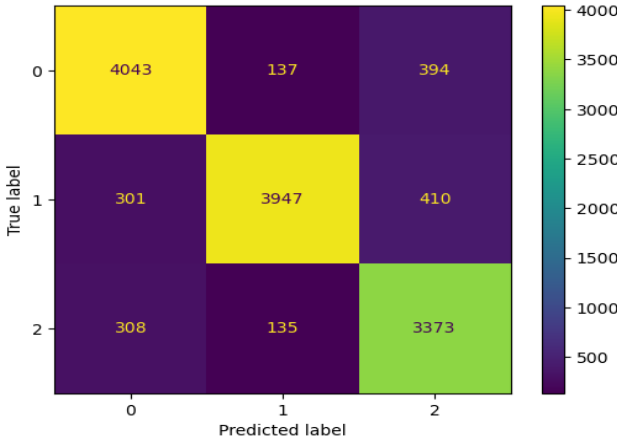

FIGURE 4. Confusion matrix of LR

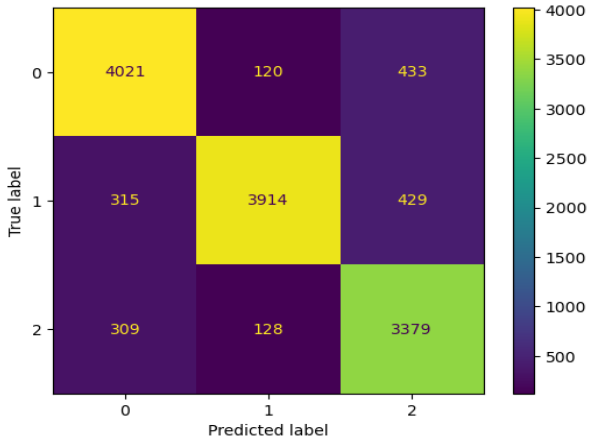

FIGURE 5. Confusion matrix of SVM

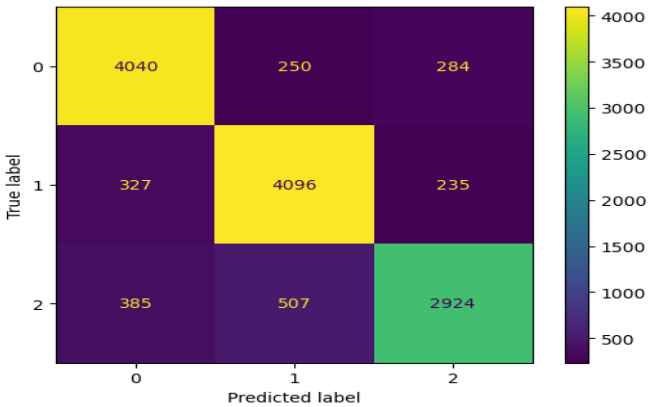

FIGURE 6. Confusion matrix of NB

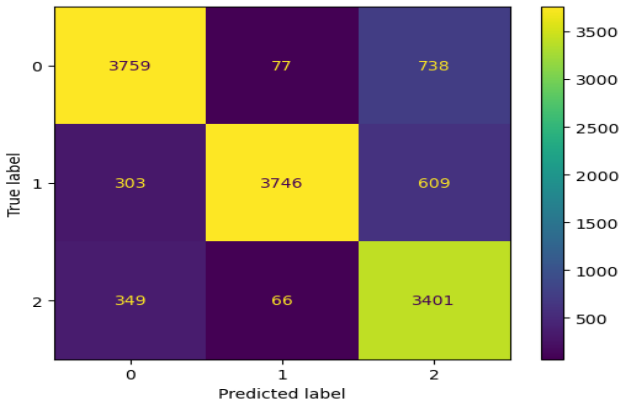

FIGURE 7. Confusion matrix of RF

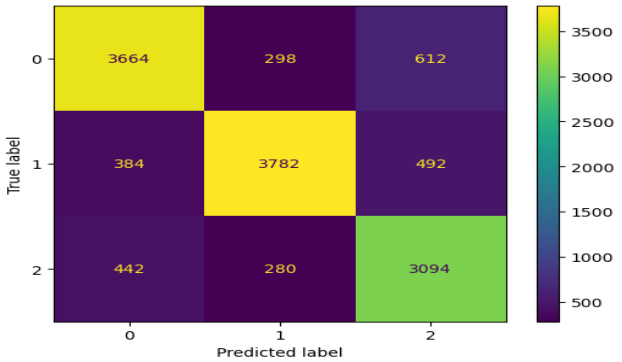

FIGURE 8. Confusion matrix of DT

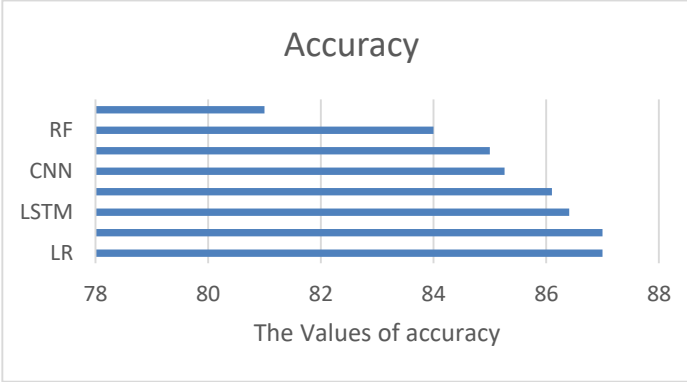

FIGURE 9. Applied models accuracy

We compared the performance of the proposed framework with the latest advanced ASA models. The accuracy values of the applied models in this work are compared with other results from various papers. Other papers used different Arabic datasets such as AJGT [31], ArTwitter dataset [32], and Arabic Sentiment Tweets Dataset (ASTD) [33]. These papers applied singular or hybrid models of different classifiers. In this work, we applied the LR classifier and SVM classifier to the Arabic Companies Reviews dataset, producing the best result of high Accuracy of 87%, and comparing with [34], it is found that the accuracy value is 82% when using the LR classifier on the AJGT dataset. It used KNN, which also produced an accuracy value of 82% if compared with other applied classifiers in this paper as accuracy values are 86.10%, 85.26%, 85%, 84%, and 81% (CNN-LSTM), (CNN), (MNB), (RF), and (DT) respectively. LSTM networks are applied in this paper and produce an accuracy value of 86.41%, but a 77.62% accuracy value was produced from the same classifier applied to the ASTD dataset. When the CNN classifier was applied to the ArTwitter dataset, it produced an accuracy value of 82.48% in [35]. Applying the CNN classifier to the ASTD dataset produced an accuracy value of 77.62%[36]. The researchers in [13], developed a hybrid system for Arabic sentiment analysis (SA) that combines lexicon-based and machine-learning approaches. They conducted experiments using different datasets, including ASTD and ArTwitter, and employed Support Vector Machine (SVM), Logistic Regression (LR), and Recurrent Neural Network (RNN) classifiers. The LR classifier achieved the best performance with an accuracy of 83.73%, followed by RNN with 81.62% and SVM with 81.52%. According to [14], the researcher applied the bi-LSTM model to the Arabic tweets and achieved accuracy scores of 0.7605 with a Macro-F1 score.

5. Conclusion and future work

In today's fast-paced world marked by technological advancements, a significant portion of the global population relies on the Internet for communication, entertainment, and various professional tasks. Given the vast data available, sentiment analysis has become a crucial process. It involves extracting reviews, tweets, and comments from diverse social media platforms to categorize their polarity as neutral, harmful, or positive. There have been limited studies addressing sentiment analysis of the Arabic language. We used TF-IDF and word embedding for textual representations. We applied five machine learning models (LR, SVM, MNB, RF, DT) and three Deep Learning models (CNN, LSTM, hybrid CNN-LSTM). The input reviews were classified into positive, neutral, and hostile. Experiments have demonstrated that the LR and SVM classifiers attained an accuracy of 87%, the highest among all classifiers. At the same time, the other classifiers (LSTM), (CNN-LSTM), (CNN), (MNB), (RF), and (DT) achieved accuracies of 86.41%, 86.10%, 85.26%, 85%, 84% and 81% respectively. There is an opportunity to use other alternatives of DL approaches or complex architectures for both classifiers of LSTM networks and CNN across several forms of datasets in the future. We can use a pre-trained model for word2vec and test the performance. Also, future studies can apply transformers (BERT) on the same dataset.

References

[1] Ahmed A. & Nouh E., (2020), "A comprehensive study for Arabic Sentiment Analysis (Challenges and Applications)", Egyptian Informatics Journal 21 (2020) 7–12

[2] Ruba O., Duha A., Esra A., and Osama H., (2021), "Arabic Aspect-Based Sentiment Analysis: A Systematic Literature Review", IEEE Access, Vol. 9, pp. 152628-152645

[3] Elnagar, A., Yagi, S., Nassif, A. B., Shahin, I., & Salloum, S. A. (2021). Sentiment analysis in dialectal Arabic: a systematic review. Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021, 407-417.

[4] Elhassan, N.; Varone, G.; Ahmed, R.; Gogate, M.; Dashtipour, K.; Almoamari, H.; El-Affendi, M.A.; Al-Tamimi, B.N.; Albalwy, F.; Hussain, A. Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning. Computers 2023, 12, 126. https://doi.org/10.3390/ computers12060126

[5] Musleh D., Alkhwaja I., Alkhwaja A., Alghamdi M., Abahussain H., Alfawaz F., Min-Allah N. and Abdulqader M., (2023), "Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation", big data and cognitive computing, 7, 127, pp 1-16: https://www.youtube.com/watch?v=dnK5lqpyEPg

[6] Samah M. Alzanin, Aqil M. Azmi, Hatim A. Aboalsamh, (2022)" Short text classification for Arabic social media tweets", Journal of King Saud University – Computer and Information Sciences
Volume 34, Issue 9, October 2022, Pages 6595-6604

[7] Lamia, M. A., Gawaher, S. H., Nasser, H. A.(2020). A survey on sentiment analysis algorithms and techniques for Arabic textual data. Fusion: Practice and Applications, 2(2), 74-87.

[8] Al-Smadi M., Qawasmeh O., Al-Ayyoub M., Jararweh Y. & Gupta B., (2018-a), "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.

[9] Ammar M. and Rania K., (2019), "Deep learning approaches for Arabic sentiment analysis", Social Network Analysis and Mining, 9:52, pp 1-13.

[10] Al-Smadi M., Talafha M., Al-Ayyoub M. & Jararweh Y. (2019), "Using long–term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews, International Journal of Machine Learning and Cybernetics Vol.10, No. 8, pp. 2163–2175.

[11] Abdulhakeem Q., Ahmed S., and Saman H., (2020), "Arabic Sentiment Analysis (ASA) Using Deep Learning Approach", Journal of Engineering, Vol. 26 No. 6, pp 85-93.

[12] Alharbi, A.; Kalkatawi, M.; Taileb, M., (2021), "Arabic Sentiment Analysis Using Deep Learning and Ensemble Methods", Arabian Journal for Science & Engineering, Vol. 46, Issue 9, pp. 8913-8923...

[13] Asma B, and Zouhour N., (2023), "Sentiment analysis classification for text in social media: application to Tunisian dialect", International Journal on Cybernetics & Informatics (IJCI) Vol. 12, No.2, pp.313-326.

[14] Ahmed B. (2023), "Improved Deep Learning Sentiment Analysis for Arabic", Journal of Theoretical and Applied Information Technology, Vol.101. No 3, pp 1251-12-60

[15] Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y. & Qawasmeh, O., (2018-b), "Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features. Information Processing and Management, Vol. 56, No. 2, pp. 308–319.

[16] Elshakankery K, and Ahmed M., "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis", Egyptian Informatics Journal, Vol. 20, No. 3, 2019, pp. 163-171.

[17] Gamal D, Alfonse M, El-Horbaty E., and Salem AB, "Twitter benchmark dataset for Arabic sentiment analysis", International Journal of Modern Education and Computer Science, Vol. 11, No. 1, 2019, pp. 33-38.

[18] Alyami, S., Olatunji, S., (2020), "Application of Support Vector Machine for Arabic Sentiment Classification Using Twitter-Based Dataset ", Journal of Information and Knowledge Management, Vol. 19, No. 1, pp. 1–13.

[19] Alsalman, H., (2020), "An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter social media. 3rd International Conference on Computer Applications and Information Security, Riyadh, Saudi Arabia, 19–21 March 2020.

[20] Alharbi, L., Qamar, A., (2021), "Arabic Sentiment Analysis of Eateries' Reviews: Qassim region Case study", 4th National Computing Colleges Conference, NCCC 2021, Taif, Saudi Arabia, 27–28 March 2021.

[21] Govindan, V., & Balakrishnan, V. (2022). A machine learning approach in analyzing the effect of hyperboles using negative sentiment tweets for sarcasm detection. Journal of King Saud University-Computer and Information Sciences, 34(8), 5110-5120.

[22] Musleh, D., Alkhales, T., Almakki, R., Alnajim, S., Almarshad, S., Alhasaniah, R., Aljameel, S. & Almuqhim, A., (2022), "Twitter Arabic sentiment analysis to detect depression using machine learning. Computers, Materials & Continua, Vol 71, No. 2, pp. 3463–3477.

[23] Maria Y. and bdulla A., (2022), "Analysis and Evaluation of Two Feature Selection Algorithms in Improving the Performance of the Sentiment Analysis Model of Arabic Tweets", IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 6, 2022, pp. 705-711.

[24] Arabic Companies Reviews For Sentiment Analysis (kaggle.com), retrieved November 2023.

[25] Socher, R., 2014. Recursive deep learning for natural language processing and computer vision Ph.D. thesis. Stanford University.

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

[27] Lilleberg, J., Zhu, Y., Zhang, Y., 2015. Support vector machines and word2vec for text classification with semantic features, in: IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC'15), pp. 136–140.

[28] Yang, Z.T., Zheng, J., 2016. Research on Chinese text classification based on Word2vec. Proceeding of the Second IEEE International Conference on Computer and Communications (ICCC), 1166–1170.

[29] Almuzaini, H.A., Azmi, A.M., 2020. Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization. IEEE Access 8, 127913– 127928.

[30] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space, in: First International Conference on Learning Representations (ICLR 2013).

[31] Alomari, K.M., ElSherif, H.M. and Shaalan, K. (2017) 'Arabic tweets sentimental analysis using machine learning', Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Proceedings, Part I, 27–30 June, Springer International Publishing, Arras, France, pp.602–610.

[32] Altowayan, A.A. and Tao, L. (2016) 'Word embeddings for Arabic sentiment analysis', 2016 IEEE International Conference on Big Data (Big Data), December, IEEE, pp.3820–3825.

[33] Nabil, M., Aly, M. and Atiya, A. (2015) 'Astd: Arabic sentiment tweets dataset', Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, September, pp.2515–2519.

[34] Bolbol, N.K. and Maghari, A.Y. (2020) 'Sentiment analysis of Arabic tweets using supervised machine learning', 2020 International Conference on Promising Electronic Technologies (ICPET), December, IEEE, pp.89–93.

[35] Dahou, A., Elaziz, M.A., Zhou, J. and Xiong, S. (2019) 'Arabic sentiment classification using convolutional neural network and differential evolution algorithm', Computational Intelligence and Neuroscience, Vol. 2019.

[36] Alayba, A.M., Palade, V., England, M. and Iqbal, R. (2018) 'A combined CNN and LSTM model for Arabic sentiment analysis', Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Proceedings, 27–30 August, Hamburg, Germany, pp.179–191, Springer International Publishing.