Systems and Soft Computing 7 (2025) 200203

Contents lists available at ScienceDirect

Systems and Soft Computing

journal homepage: www.journals.elsevier.com/soft-computing-letters

A systematic assessment of sentiment analysis models on iraqi dialect-based texts

Hafedh Hameed Hussein, Amir Lakizadeh *

Department of Computer Engineering and Information Technology, University of Qom, Qom, Iran

ARTICLE INFO

Keywords:
Sentiment analysis
Iraqi dialect
Deep learning
Polarity classification

ABSTRACT

Social media allows individuals, groups, and companies to openly express their opinions, creating a rich resource for trend assessments through sentiment analysis. Sentiment Analysis (SA) uses natural language processing (NLP) to interpret these opinions from text. However, Arabic sentiment analysis faces challenges due to dialect variations, limited resources, and hidden sentiment words. This study proposes hybrid models combining Convolutional Neural Networks with Long Short-Term Memory called as CNN-LSTM, CNN with Gated Recurrent Unit called as CNN-GRU. and AraBERT, a deep transformer model, to enhance Iraqi sentiment analysis. These models were evaluated against various machine learning and deep learning models. For feature extraction, we utilized Continuous Bag of Words (CBOW) for deep learning models and BERT for the AraBERT model, while TF-IDF was used for machine learning models. According to the experimental results, the AraBERT model has been able to achieve superior performance and significantly improve the accuracy of sentiment analysis in case of Iraqi dialect-based texts.

Introduction

Within Natural Language Processing (NLP), Sentiment Analysis (SA) is a specialist field dedicated to the recognition and understanding of the emotions or points of view voiced in a given text. Many people use several online sites to express their ideas and opinions [1]. Examining user-generated data is thus crucial to monitor public mood and support decision-making. Dealing with Sentiment Analysis (SA) presents a number of difficulties including informal writing styles and language-specific problems that must be overcome. Moreover, many languages feature a lot of words with different connotations and meanings. This limits the tools and resources available for every language [2] .

Arabic ranks sixth among all the languages spoken worldwide, with a billion or so speakers. Its special traits make general natural language processing (NLP) challenging. Arabic Sentiment Analysis (ASA) finds particular difficulties since the language has a different structure than other languages. Arabic language has three main variants apart from its unique structure: Arabic Dialect (AD), Standard Arabic (MSA), and Classical Arabic (CA). Official and educational correspondence uses MSA; CA is used in literary and religious contexts. AD lacks a consistent spelling though, and mostly reflects the spoken Arabic. It consists of five main groups: Egyptian, Levantine, Gulf, Iraqi, and Maghrebi [3]. The most often used form of communication in daily contacts and media in Iraq under the category of Iraqi dialect is Iraqi Dialect (ID) [4].

Sentiment analysis for Iraqi dialects has several interesting challenges and constraints, mostly related to the lack of publicly available, suitably labeled datasets created especially for this use. Using many sources including Facebook, Twitter, and movie reviews, Tabel (1) summarizes recent studies on sentiment analysis for Iraqi dialect datasets. These datasets range in scale from hundreds of samples to smaller collections of tweets, thus they differ greatly. Usually, labels describe positive, negative, and sometimes neutral attitudes. Naïve Bayes (NB), support vector machines (SVM), decision trees (DT), K-nearest neighbors (KNN), and Rough Set Theory (RST) have among other machine learning approaches been used. In particular cases, some studies attained great accuracy; RST reached 94 % and SVM attained over 92 %. Furthermore, investigated for sentiment classification are clustering methods including mean shift, DBSCAN, and K-means. These initiatives show how to address sentiment analysis issues for the Iraqi dialect by using both conventional and modern approaches.

Aiming at benchmarking text sentiment analysis and polarity classification for Iraqi texts, we present previous research [5] introduces a newly compiled corpus of Facebook comments in Iraqi dialect (called

* Corresponding author.
E-mail address: lakizadeh@qom.ac.ir (A. Lakizadeh).

https://doi.org/10.1016/j.sasc.2025.200203
Received 27 October 2024; Received in revised form 17 December 2024; Accepted 6 February 2025
Available online 7 February 2025

IRAQISAT), so addressing the issue of Iraqi Dialect Dataset. From Facebook comment data, the Iraqi corpus was generated by Facepager program. Within the framework of the Arabic language, sentiment analysis research benefits much from this corpus.

In this work we investigate sentiment analysis approaches applied to Iraqi dialect datasets in their whole using a range of machine learning, deep learning, and transformer-based methods. For conventional machine learning we use algorithms including Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbor (KNN). We also investigate novel deep learning architectures including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and hybrid models CNN-GRU and CNN-LSTM in order to reflect the contextual subtleties of the text.

We use the transformer-based AraBERT model—pre-trained especially for Arabic text—to improve sentiment analysis performance even more. We want to enhance the understanding of dialectal Arabic and get better classification results by fine-tuning AraBERT on our dataset, which comprises of data gathered from public Iraqi Facebook sites. This multifarious approach enables us to assess and contrast, in sentiment analysis for Iraqi dialects, the efficiency of conventional and modern approaches.

The paper is structured as follows: Section 2 offers a summary of related works in Arabic sentiment analysis together with a discussion of several methods spanning machine learning to transformer-based approaches. Section 3 describes the suggested approach together with data preparation, feature extraction, and used models. Section 4 shows the experimental results, so highlighting the performance of the applied methods. The results are covered in Section 5 together with a thorough study and interpretation of them. Section 6 marks the end of the research and summarizes the main findings together with recommendations for next directions.

Related works

Sentiment analysis (SA) has been studied in many languages using a range of techniques quite extensively. Because of the linguistic challenges Arabic sentiment analysis presents—its rich morphology, varied dialects, and informal writing style common on social media platforms—it has drawn a lot of attention lately. This section summarizes significant research efforts in Arabic sentiment analysis and classifies them into three main categories based on the applied techniques—machine learning, deep learning, and transformer-based models like AraBERT, especially designed for Arabic text. By raising the accuracy and dependability of sentiment classification in Arabic datasets, these techniques have tremendously advanced the discipline (Table 1).

Machine learning methods

Sentiment analysis has become increasingly popular recently, especially with relation to Arabic dialects—which have not gotten much attention in previous studies. In studies conducted by [7], 1080 Facebook comments and posts from the Iraqi dialect were examined in a dataset. The dataset consisted in 540 entries in every one of the two sentiment categories: positive and negative. Using many machine learning models including NB, SVM, LR, DT, RF, and KNN, the researchers examined attitudes. Out of all the classifiers, the Naïve Bayes one proved better with an F-measure of 82 % and an accuracy rate of 81 %. On the other hand, the KNN model registered with the lowest accuracy—57 %). These results underline the need of more research on sentiment analysis techniques especially meant for Arabic dialects since the complexity of informal language use on social media platforms highlights the effectiveness of Naïve Bayes in analyzing dialectal emotions.

Deep learning methods

Seeking to enhance Arabic sentiment analysis, (Saleh et al., 2022) present in their work an ideal heterogeneous ensemble deep learning model. Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) are three pre-trained deep learning models stacked in this work with meta-learners including Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). Convolutional neural networks (CNNs) help to capture local features and patterns in the text data, so allowing n-grams and significant pattern recognition for sentiment classification. Retaining knowledge over long sequences helps LSTMs solve the vanishing gradient problem usually present in conventional RNNs. This combination allows one to investigate the temporal dependencies in the textual data more powerfully. Using KerasTuner and grid search helps the ensemble model to be optimal, so improving performance. The model is evaluated with accuracy, precision, recall, and F1-score measures on three benchmark Arabic datasets: ASTD (Arabic Sentiment Tweets Dataset), ArSenTD-Lev (Levantine Dialect Arabic Sentiment Twitter Dataset), and Hotel Arabic Reviews Dataset. By its performance above traditional machine learning techniques, the proposed ensemble model shows how well it manages the complex morphological structures and several dialects of the Arabic language. The ensemble model especially achieved higher accuracy and F1-scores across all datasets, compared to other methods including Decision Tree (DT), K-Nearest Neighbor (KNN), and Naive Bayes (NB): ASTD at 93.1 %; ArSenTD-Lev at 94.5 %; the Hotel Arabic Reviews Dataset at 95.2 %. CNNs and LSTMs taken together significantly enhance sentiment analysis performance, so proving the model's ability to control the intricate subtleties of Arabic text data.

Using both machine learning and deep learning approaches, (Alshutayri et al., 2022) examine sentiment analysis for Arabic tweets. The dataset consists of 32,186 tweets from many Arabic dialects—including Modern Standard Arabic (MSA) and dialects including Levantine, Egyptian, and Gulf Arabic. There abound in these tweets

Tabel 1
A Comparative analysis for studies worked on Iraqi Dialects.

Ref.	Dataset size	Data Source	Classes	Used Classifiers	Best classifiers	Accuracy
[6]	1200	Politic Facebook pages	Pos, Neg	Ad boost, Multinomial-NB KNN	KNN	80 % Acc.
[7]	1080	Facebook	Pos, Neg	NB, SVM, LR, DT, RF, and KNN	NB	81 % Acc.
[8]	5000	Facebook	Pos, Neg, Neu, Spam	Build Dataset		
[9]	4000	Facebook	Pos, Neg	LR, DT, SVM, and NB	SVM	82 % Acc.
[10]	1170	Tweeter	Pos, Neg, IDK	SMO SVM, Lib SVM (Sequential minimum optimization)	SVM	78 % Acc.
[11]	hundreds of tweets	Tweeter	Pos, Neg, Neu	Lexicon based algorithm with (EM, DBSCAN, K-means, mean shift, and agglomerative clustering techniques)	K-means	72 % Acc.
[12]	Dataset of [8]	5000 Facebook	Pos, Neg, Neu, Spam	CNN-GRU	CNN-GRU	92.4 F1
[12]	Dataset of [9]	4000 Facebook	Pos, Neg	CNN-GRU	CNN-GRU	92.5 F1

positive, negative, and neutral attitudes. Among the several methods of machine learning applied were logistic regression (LR), support vector machine (SVM), and multinomial and Gaussian Naive Bayes. Of these, SVM came in with the best accuracy 63 %. The work also included deep learning methods; the Long Short-Term Memory (LSTM) network raises accuracy to 70 %. Although the dataset was imbalanced22,275 neutral tweets against 4920 positive and 4806 negative tweets—some sentiment labels were false. These components presented challenges for additional performance improvement of the model and resulted in rather low accuracy.

(Al-Hassan & Al-Dossari, 2021) investigate deep learning approaches for Arabic Twitter hate speech identification. Comprising five separate classes—none, religious, racial, sexism, or general hate—they assembled an 11,000-tweet dataset. Reflecting the variety of the Arabic-speaking community, the tweets in many Arabic dialects reveal The researchers compared four deep learning models—LSTM, CNN+LSTM, GRU, and CNN+GRU against a baseline Support Vector Machine (SVM) model.By means of a Convolutional Neural Network (CNN), the CNN+LSTM model essentially controls spatial hierarchies and detects n-grams first to capture local features and patterns inside the text data. These characteristics are then fed into a Long Short-Term Memory (LSTM) network, which handles the issue of vanishing gradients and stores information over long sequences so managing the problem of temporal dependencies. Similar local feature generation from a CNN using which a Gated Recurrent Unit (GRU) network is then fed generates the CNN+GRU model. Even if their simplified architecture helps to lower computational complexity while still enabling to manage long-term dependencies in the text, GRUs are comparable to LSTMs. With an average recall of 75 %, deep learning models beat SVM; CNN+LSTM model performed best with 72 % precision, 75 % recall, and 73 % F1-score. This work addresses the difficulties given by the complexity and dialectical variations of Arabic tweets by showing the efficiency of merging CNNs with LSTMs and GRUs for hate speech detection.

(Al-Bayati et al., 2020) explore Arabic Sentiment Analysis (ASA) using a deep learning approach. By means of Long Short-Term Memory (LSTM) networks augmented with Arabic word embeddings, they capture the rich morphological structure of the language. Comprising 47,000 training and 11,000 testing, the dataset consists of 58,000 tweets penned in Modern Standard Arabic (MSA) as well as several dialects including Levantine, Egyptian, and Gulf Arabic. Sequentially addressing these embeddings, the LSTM model captures the sentiment and context of every tweet. Training the LSTM network with backpropagation across time (BPTT) helps to maximize the weights and minimize prediction error. The results show an accuracy of roughly 82 %, hence underlining how effectively deep learning techniques control the complexity of Arabic sentiment analysis. This approach highlights how well LSTM networks could address issues resulting from the several dialects and complex Arabic language morphological structures.

Using both machine learning and deep learning techniques, (Onan, 2019) investigates sentiment analysis (SA) of teacher assessment reviews. Comprising 154,000 reviews taken from well-known teacher review sites, the dataset is either positive or negative. Together with several ensemble learning approaches to improve classification performance, the author uses five machine learning algorithms: Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Random Forest (RF). Apart from these approaches, Onan makes use of five deep learning algorithms: Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), CNN, RNN-AM with bidirectional RNN with attention mechanism (RNN-AM), and RNN-AM with CNN Every deep learning model is matched with unique word embeddings to successfully convey the textual semantic meaning. Among these, using the RNN with GloVe word embeddings produced the best degree of accuracy 98.29 %.

AraBERT transformer

AraBERT is the first Arab transformer kind. They accomplished this using Google's developed BERT transformer model. Wide spectrum of current standard Arabic and other dialects was used for training the AraBERT model. After that, it was tested on several projects including SA, Named Entity Recognition, and Question Answering. For Arabic NLP [13], this has benefited the discipline. Arabert is especially meant to solve the difficulties with processing the Arabic language, including the presence of several negations, different accents, and difficult words. For companies and companies looking for understanding of the ideas and emotions of their customers, Arabic script becomes helpful since it can grasp its context and semantics. Extremely helpful for Arabic natural language processing (NLP), this tool has shown great performance in many Arabic NLP projects. BERT and Arabert have lately been used in several NLP and SA studies. [14] developed their Arabic BERT method on the BERT model in order to solve problems arising with the Arabic language. The researchers conducted a comparative analysis with earlier studies using traditional machine learning and deep learning approaches on the same datasets using the method on five different datasets. Using a variety of machine learning approaches, [15] applied AraBERT on several sets of data from many areas, including movie reviews, restaurant reviews, and product reviews. Examining comments and reviews from many websites, the authors Moubtahij [16] used the ARev dataset—40,000 Arabic reviews with both positive and negative emotions. They used the AraBERT model to do this. Their method worked 92.5 % of the time.

The proposed methodology

This section provides a justification of the dataset used in our research, the methods of preparation and preprocessing implemented to the Iraqi sentiment analysis dataset, and the feature extracting methods applied. It also explains the several techniques applied: machine learning approaches, deep learning architectures, hybrid models, and the AraBERT transformer model. To increase sentiment analysis performance, we especially use the pre-trained AraBERT model—which has been tuned for the Arabic language. Moreover, discussed are the performance evaluation methods—accuracy, precision, recall, F-measure—used in these models. Fig. 1 highlights the data preparation through to model evaluation process and shows the recommended approach.

Dataset

The study's objective is to implement different analysis models to our created IRAQIDSAD corpus [5]. The created corpus contains 14,141 annotated comments for sentiment analysis of Iraqi dialects. It was collected from four Iraqi Facebook common sites ("دلیل مطاعم بغداد": Baghdad Restaurants Directory); ("برنامج ولاية بطيخ": Melon City show); ("ستيفن نبيل": Steven Nabil) and ("بغداد": Baghdad) page. The corpus includes three classes: positive, negative, and normal had the same comment size as shown in Fig. 2.

Data preprocessing

Given the presence of noise in the Iraqi dialect on Facebook social networks, it is necessary to preprocess the data in order to decrease the extensive vocabulary. This reduced vocabulary will then be utilized as a feature in the future. The six sub-stages included in the pre-processing data stage are explained in detail below.

- Filtering: Following the data collecting phase, we have acquired a substantial number of Facebook comments. However, we have identified several issues with the obtained remarks that undermine their correctness. Comments that don't have any opinion (comments that contain just one character or simple (e.g., ".", " م").)

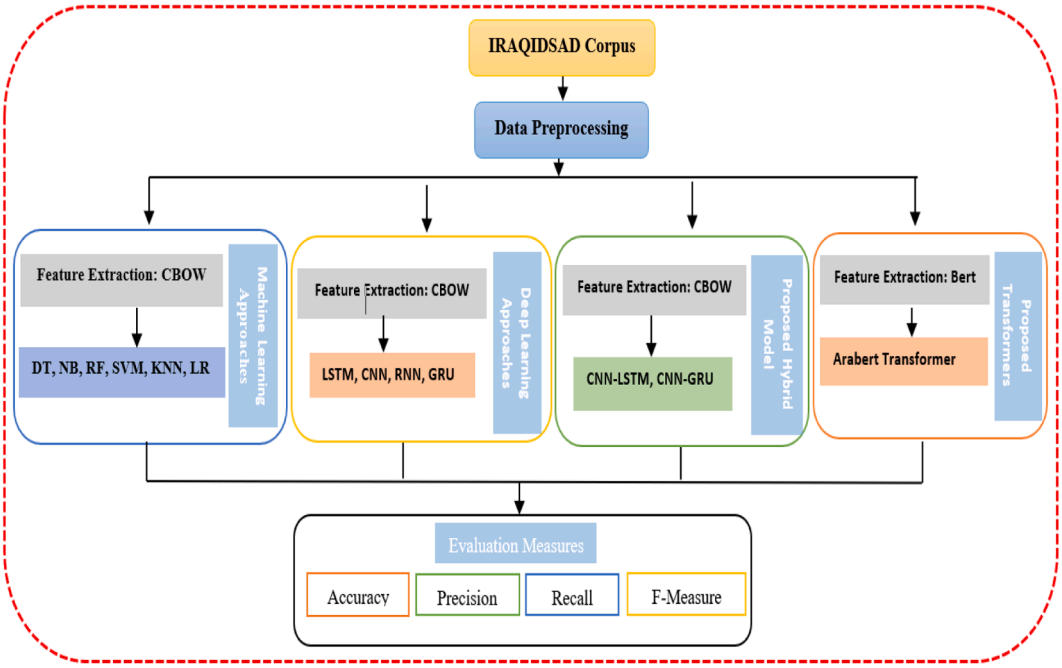


Fig. 1. Block diagram of the proposed model.

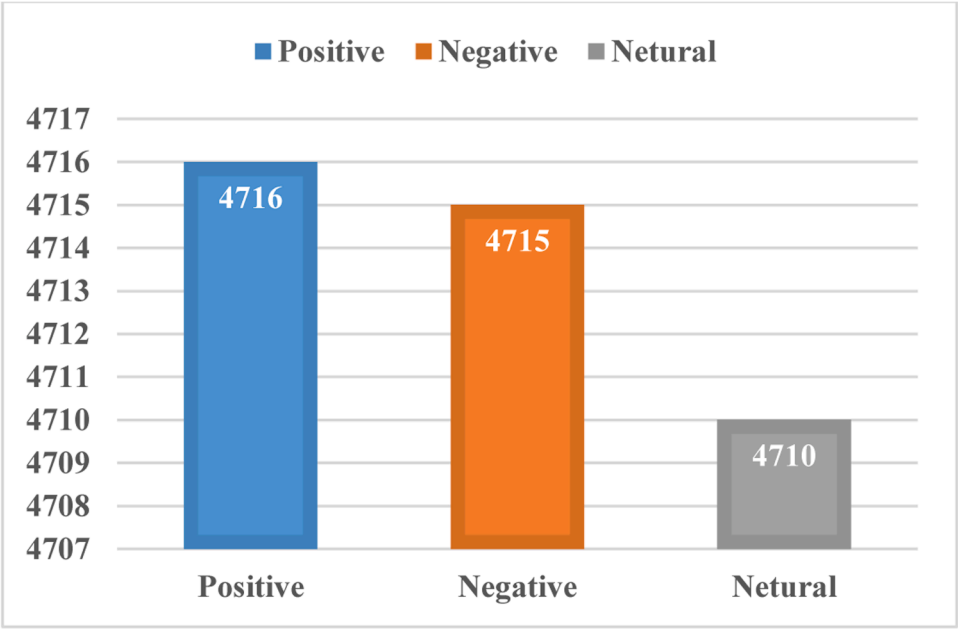


Fig. 2. Distribution of Positive, Negative and Neutral labels in the dataset.

- Comments with serious bad words cannot be acceptable in any way.
- Comments are written in other languages (truckman, Kurdi, English).
- Comment with Facebook reactions (like love, haha, wow, sad, angry).
- Comments that contain just a tagged name.
- Comments on redundancies.
- Links, remarks, and images (referred to as 'photo scraps' or '[[photo]]') are included in the comments. It is necessary to delete all of these comments.

- **Annotation:** Following the sub-stage filtration, it will be necessary to manually assign labels to the comments, categorizing them into two groups using (0,1,2) labels. These labels correspond to the positive, negative, and normal classes, respectively. We will carefully analyze every remark and categorize it based on the assumption that each comment expresses a viewpoint. After that, we will need to display the labeled data to some experts to show their opinion about the annotation of each comment.
- **Tokenization:** Initially, the text will be divided into individual tokens or words, allowing for distinct treatment of each token. The tokenization process is based on the presence of gaps between tokens.
- **Data Cleaning**: Cleaning the data include the following steps:
  - Remove special char such as "@, $, %, ^, &, *,: < >" and so on.
  - Remove numbers and non-Arabic text.

- Remove the repeated char form token and exclude the word that has repeated char in its original form.
- Normalize the letter as follows:
- (أ ,آ ,إ) replaced with (ا)
- (ة) replaced with (ه) from the end of the word.
- (ؤ) replaced with (و).
- (ظ) replaced with (ض)

- **Stemming**: refers to the process of reducing words to their original root form. In our work, this sub-stage presents a challenge as we are dealing with dialect rather than modern standard Arabic. Obtaining the pattern from the Iraqi dialect is difficult, which adds to the complexity. Furthermore, since we are working with data from social media where users write comments and posts in various ways, it is challenging to establish rules for stemming words. While we will attempt to reduce words as much as possible, achieving the exact stem is nearly impossible. Additionally, it is important to note that there are regional variations in dialect within Iraq.
- **Stop Word Removal:** In the field of natural language processing, stop words refer to words that lack semantic significance, such as pronouns, prepositions, and conjunctions. These words are thus excluded prior to categorization. Stop words vary based on language and dialect, hence there is no pre-established compilation of stop words; this investigation will construct a compilation of stop words.

Feature extraction approaches

Feature extraction is one of the most important actions that must be taken with arbitrary data such as text or images in order to convert it into digital data so that it can be used for machine learning, Deep learning, and Transformer approaches [17,18]. In this study, Term frequency-inverse document frequency (**TF-IDF**) will be used for Machine learning, **CBOW** for Deep Learning, and **Bert** for Arabert transformer.

TF-IDF is a statistical method used to assess the relevance of a word to a document in a collection of documents. In order to do this, the frequency of a word in a single document is multiplied by its inverse document frequency over a collection of documents. Its applications are many, with the primary focus being automated text analysis. It is especially useful in evaluating the effectiveness of words within Natural Language Processing (NLP) based machine learning systems. TF-IDF became a framework for document search and information retrieval. The algorithm operates by increasing proportionally to the frequency of a word's occurrence in a document, with the number of papers in which the term appears acting as a balancing factor. Refer to reference [19] for the Term Frequency-Inverse Document Frequency (TF-IDF) computation formula.

CBOW is a widely used algorithm applied in deep learning and natural language processing. Unlike traditional models that predict a word given its context (like Skip-gram), CBOW predicts the target word based on its surrounding context words. This approach is particularly efficient for feature extraction in scenarios where semantic meaning and context play crucial roles, such as in sentiment analysis or language modeling tasks. CBOW works by training a neural network to learn the distributional properties of words based on their contexts, thus capturing syntactic and semantic relationships within a corpus of text. By embedding words into a continuous vector space, CBOW enables downstream tasks in deep learning models to leverage these embeddings for improved accuracy and efficiency [19].

Machine learning methods

This section will provide a brief explanation of the different machine learning techniques used in this study. Every one of these techniques is used on the feature-extracted data to assess their sentiment classification efficacy. These machine learning approaches offer a framework for dataset analysis and act as standards for comparison with more sophisticated approaches including deep learning and transformer-based models.

A. Random Forest (RF)

A popular ML algorithm called RF mixes the output of various decision trees to get a single outcome. As it addresses classification and regression issues, its adoption results from its versatility and ease of use [20].

B. Decision Tree (DT)

DT is a type of flowchart-like tree structure where each leaf node represents the result, and each inner node represents a feature (or attribute). It is a member of the supervised learning algorithm family [21]. It is a picture that accurately mimics the level of human thought. Because of this, decision trees are simple to comprehend and interpret. It is used for solving problems of regression and classification. The primary algorithm for creating decision trees is ID3. Entropy and Information Gain are used by ID3 to generate decision trees [22].

C. Support Vector Machine (SVM)

Classification issues can be solved using the SVM method of supervised ML [23]. Each data point is represented using the SVM algorithm as a point in an n-dimensional space (n is the number of features), with the value of each feature being the value of a specific coordinate. After then, the classification was finished by determining the hyper-plane that best separates the two classes. SVM has three kernels [24].

D. Naïve Bayes (NB)

The classification algorithm NB is appropriate for classification problem. It is a supervised classification method that assigns class labels to instances or records using conditional probability to categories future items. Defined by the formula [25,26].

E. K-Nearest Neighbors' algorithm (KNN)

One of the most fundamental yet crucial categorization methods in ML is KNN. It falls under the category of supervised learning and has numerous applications in data mining, intrusion detection, and pattern recognition. Where data is grouped in order to calculate the likelihood that a given data point will belong to a particular group depending on the group to which the data points nearest to it belong. It used one from those three distances [27]. The family of supervised machine learning models includes logistic regression [28]. It is employed in the classification and prediction of data. Based on a collection of independent variables, logistic regression calculates the likelihood of an event occurring, such as voting or not voting. The dependent variable is confined between 0 and 1 because the outcome is a probability.

F. Logistic Regression (LR)

A logit transformation **(**LR**)** is performed to the odds in logistic regression, which is the probability of success divided by the probability of failure. The logistic function is represented by the following formulas [29].

Deep learning approach

We introduce the several deep learning techniques used in this work for sentiment analysis in this section. Using modern neural network architectures, these techniques are meant to capture the intricate patterns and contextual subtleties of textual data:

A. Long Short-Term Memory-networks (LSTM)

LSTM can memorize and learn long-term dependencies, thus it is rather RNN. They often remember events from the past for protracted lengths of time. LSTMs are good in time series forecasting since they preserve information over time and recall past inputs. Four interacting layers of LSTMs are coupled in a chain-like configuration to transmit in an original way. LSTM is also used in the synthesis of voice recognition, drugs, and music [30].

B. Gated Recurrent Unit (GRU)

Though lacking an output gate, GRU follows design ideas much as a long short-term memory (LSTM). A gated recurrent unit (GRU) forms the gating mechanism of recurrent neural networks (RNN). This is a quite effective approach for the vanishing gradient issue of recurrent neural networks. Moreover, on smaller datasets [31] it beats LSTM.

C. Convolutional Neural Network (CNN)

Designed as a deep learning network, CNN [32] learns straight from data, so eliminating the need for hand-made feature extraction. CNNs are especially helpful for image pattern-seeking object, face, and scene recognition. Additionally helpful for classification of non-image data including signal, time series, and audio is they Applications requiring object recognition and computer vision abound in CNNs; examples include facial recognition and self-driving cars. A CNN consists in an input layer, an output layer, and several hidden layers in between [33].

D. Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM)

For sentiment analysis the CNN-LSTM model combines Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs). The process starts with converting input text into a word matrix, which is then local feature extracting convolutionally layered. A pooling layer then reduces these properties in dimensionality, so improving computational efficiency. Including the pooled features into an LSTM layer aids in the text's contextual information and long-term dependency capture. Following LSTM layer flattening, dense layer passing, and lastly sentiment prediction generation—that is, text classification as positive, negative, or neutral—the output layer generates as seen in Fig. 3.

E. Convolutional Neural Network with Gated Recurrent Unit (CNN-GRU)

Fig. 3 also shows the CNN-GRU model for sentiment analysis—which combines Gated Recurrent Units (GRUs) with CNNs. Input text is first converted into a word matrix and then extracted by a convolutional layer, just as with the CNN-LSTM model. A pooling layer reduces the dimensionality of these features. Once the final feature maps are obtained, a GRU layer effectively compiles contextual data and temporal dependencies. After flattening and passing the output of the GRU layer over a dense layer, the last output layer creates sentiment predictions, so classifying the text as either positive, negative, or neutral.

The arabert transformer

Made to manage natural language processing (NLP) tasks, particularly for the Arabic language, Arabert is a complex language model developed by OpenAI. Built with several transformer layers and with a bidirectional approach—that is, considering both the prior and following context of every word—the model [16] better understands its meaning. AraBERt uses its last hidden layer in sentiment analysis to forecast text sentiment. The text is first passed through a linear layer then a softmax function generates a probability distribution across specified sentiment categories. Developed at the American University of Beirut, the "aubmindlab/bert-base-arabertv02″ variation of AraBERT Designed on a BERT-based architecture with 12 layers and 110 million parameters, it has been trained on vast Arabic datasets including Wikipedia and news items. By strengthening the tokenization process, this version, Arabert v0.2, surpasses previous models in handling Arabic's complicated morphology and dialectal variations. Showing good performance on Arabic NLP benchmarks, it can be tuned for chores including text classification, sentiment analysis, and named entity recognition. In this work, bert-base-arabertv02 will be optimized to operate with an Iraqi Arabic corpus, so adjusting the model's capacity to manage dialect-specific difficulties in Iraqi Arabic for sentiment analysis. By means of its pre-trained layers, the model generates contextualized embeddings reflecting the semantic and syntactic information of the Arabic text. After that, these embeddings are optimized on our sentiment analysis dataset so that the model may learn characteristics unique to tasks and raise classification accuracy.

Results

On a given dataset, this part offers the performance comparison between several machine learning (ML) and deep learning (DL) models.

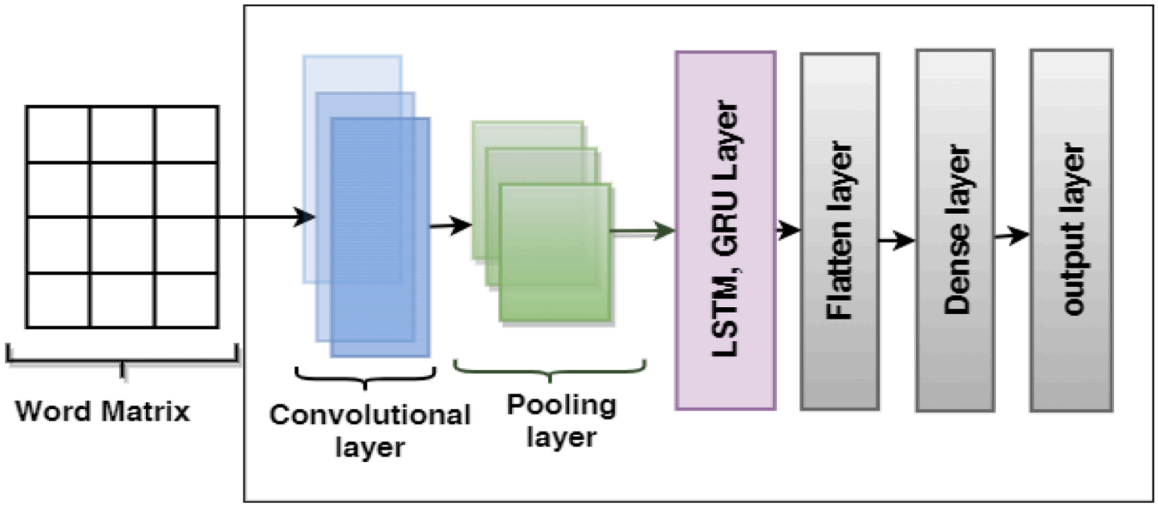


Fig. 3. Architectural design of hybrid models combining LSTM and GRU with CNN.

Two halves made of the dataset were used for training with 75 % and validation with 25 % respectively. We included the AraBERT model, four deep learning classifiers, and six machine learning classifiers. These were applied with the AraBERT library and the scikit-learn toolkit for machine learning.Precision (PRE), recall (REC), F1-score (F1), and accuracy (ACC) are among the evaluation measures applied.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

TP stands for the actual positive cases; FP stands for the false positive cases; TN stands for the actual negative cases; and FN stands for the false negative cases. Moreover, the Receiver Operating Characteristic (ROC) curve helped to evaluate the models. Graphing the True Positive Rate (TPR) and False Positive Rate (FPR), which are computed using Eqs. (5) and (6), the ROC curve visually shows the performance of the classification model at different classification thresholds. Table 2 summarizes the results.

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

TF-IDF matrix was used for evaluation of the ML models. Among the ML models, the Logistic Regression (LR) model ranked highest in accuracy (74.82 %), precision (75.14 %), recall (74.82 %), and F1-score (74.89 %). With an accuracy of 74.96 %, precision of 75.26 %, recall of 74.96 %, and F1-score of 75.03 %, the Support Vector Machine (SVM) likewise performed rather well. With an accuracy of 50.78 %, precision of 63.99 %, recall of 50.78 %, and F1-score of 47.44 %, the K-Nearest Neighbors (KNN) model performed lowest.

The Continuous Bag of Words (CBOW) matrix was the foundation for the DL models. With an accuracy of 74.34 %, precision of 76.22 %, recall of 74.34 %, and F1-score of 74.41 %, the CNN-LSTM model topped other DL models. Closely trailing with an accuracy of 74.21 %, precision of 74.22 %, recall of 74.21 %, and F1-score of 74.17 %, the GRU model With an accuracy of 68.26 %, precision of 68.24 %, recall of 68.26 %, and F1-score of 68.05 %, the LSTM model performed lowest among DL models.

The suggested model, AraBERT, which utilized the Bert matrix, demonstrated superior performance compared to both ML and DL models. It achieved an accuracy of 90.18 %, precision of 90.19 %, recall of 90.18 %, and F1-score of 90.17 %, indicating a significant improvement in all metrics. In summary, the Arabert ensemble model significantly outperforms traditional ML and DL models, establishing its effectiveness for the task at hand. Fig. 4 presents the ROC curves for each model across three sentiment classes: Positive, Negative, and Normal.

Table 2
Performance Comparison of Different Models.

Approach models	Models	Feature Extraction	Performance			
			ACC	PRE	REC	F1
ML models	RF	TF-IDF	72.64	73.72	72.64	72.66
	LR	TF-IDF	74.82	75.14	74.82	74.89
	DT	TF-IDF	69.05	71.00	69.05	68.86
	SVM	TF-IDF	74.96	75.26	74.96	75.03
	NB	TF-IDF	71.40	72.05	71.40	71.18
	KNN	TF-IDF	50.78	63.99	50.78	47.44
DL models	LSTM	CBOW	68.26	68.24	68.26	68.05
	GRU	CBOW	74.21	74.22	74.21	74.17
	CNN-LSTM	CBOW	74.34	76.22	74.34	74.41
	CNN-GRU	CBOW	72.08	72.46	72.08	72.07
Transformer model	AraBERT	Bert	90.18	90.19	90.18	90.17

Discussion

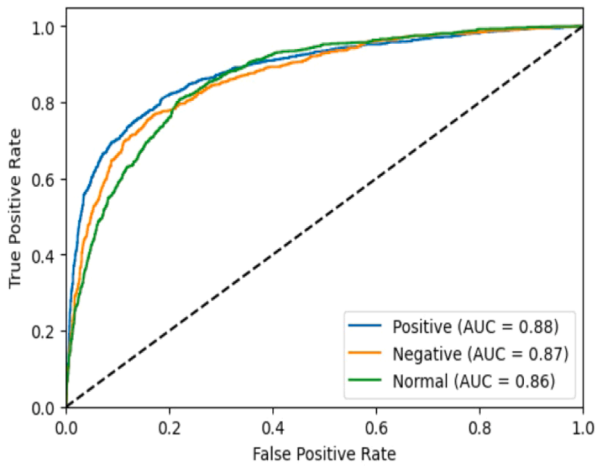
The results of our study indicate significant insights into the effectiveness of various machine learning (ML) and deep learning (DL) models in performing sentiment analysis on the Iraqi dialect dataset.

Among the ML models, Logistic Regression (LR) emerged as the most effective, achieving the highest scores across all metrics (accuracy, precision, recall, and F1-score). This suggests that LR can effectively capture the underlying patterns in the TF-IDF matrix, making it a strong baseline for sentiment analysis tasks. Support Vector Machine (SVM) also demonstrated commendable performance, closely following LR in all metrics. On the other hand, the K-Nearest Neighbors (KNN) model showed the least effectiveness, with notably lower scores. This might be due to the high dimensionality and sparse nature of the TF-IDF matrix, which KNN struggles with. In the DL category, the CNN-LSTM model achieved the highest performance, particularly excelling in precision. This indicates that combining convolutional layers with LSTM can effectively capture both local features and long-term dependencies in the CBOW matrix. The GRU model also performed well, demonstrating that it can be a viable alternative to LSTM due to its simpler architecture and faster training time. However, the LSTM model on its own did not perform as well, indicating that additional layers or architectures (like CNN) might be necessary to boost its performance in sentiment analysis tasks.

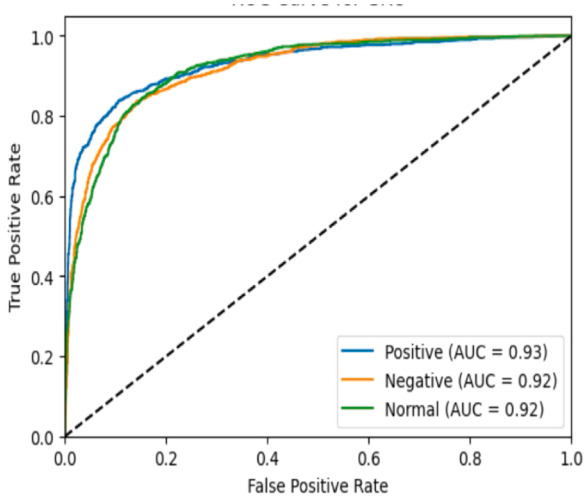
The Arabert ensemble model, utilizing the Bert matrix, outperformed all other models by a significant margin. Its high scores across all metrics highlight its superior capability in understanding and interpreting the nuances of the Iraqi dialect. This performance boost can be attributed to Bert's contextual embeddings, which provide a deeper and more nuanced representation of words based on their context within the text. The success of Arabert underscores the importance of leveraging advanced pre-trained models and fine-tuning them on specific dialect datasets to achieve optimal performance.

Conclusion

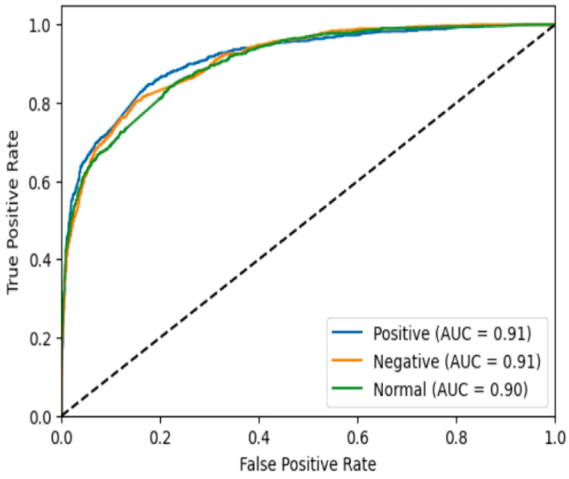
This study evaluated the performance of various machine learning (ML) and deep learning (DL) models for sentiment analysis on an Iraqi dialect dataset, offering valuable insights into their effectiveness. Among the ML models, Logistic Regression (LR) and Support Vector Machine (SVM) emerged as top performers, effectively utilizing the TF-IDF matrix to discern patterns in the text. In contrast, K-Nearest Neighbors (KNN) struggled with the high-dimensional and sparse feature space, resulting in poor performance. In the realm of DL models, the CNN-LSTM model demonstrated superior capability by combining convolutional and recurrent layers to capture both local and sequential features. The GRU model also performed well, while the standalone LSTM model lagged, highlighting the necessity for architectural enhancements. Notably, the proposed Arabert model, leveraging the Bert matrix, outshone all other models, achieving the highest scores across accuracy, precision, recall, and F1 metrics. This underscores the exceptional ability of transformer-based models to understand and interpret the nuances of the Iraqi dialect, emphasizing the importance of contextual embeddings and fine-tuning pre-trained models on specific datasets. The comparative analysis revealed that DL models generally outperformed ML models, despite their higher computational complexity and training time.



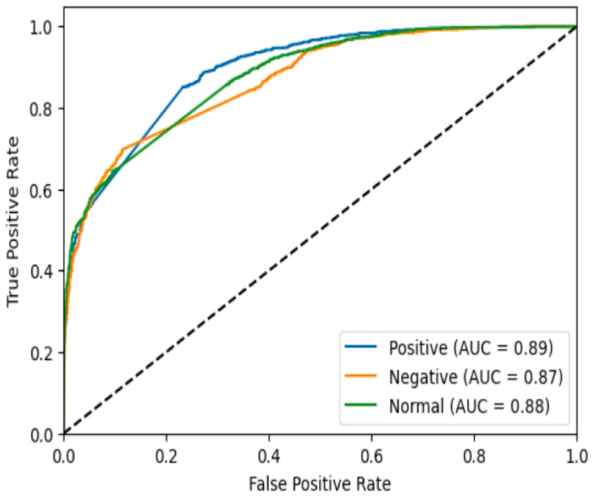
a) LSTM



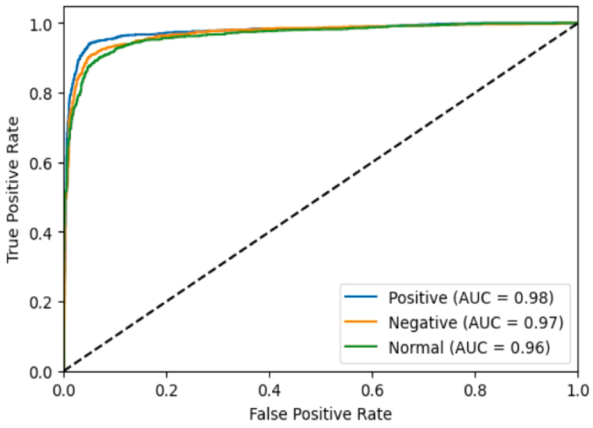
b) GRU



c) LSTM-CNN



d) GRU-CNN



e) AraBERT.

Fig. 4. ROC Curve analysis for the balanced dataset.

CRediT authorship contribution statement

Hafedh Hameed Hussein: Writing – original draft, Formal analysis. **Amir Lakizadeh:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] A. Lakizadeh, E. Moradizadeh, Text sentiment classification based on separate embedding of aspect and context, Technol. J. Artifi Intell. Data Mining 10 (1) (2022) 139–149, https://doi.org/10.22044/JADM.2021.11022.2249.

[2] M. Wnkhade, A. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, Artif Intell Rev 55 (Jul. 2022) 1–50, https://doi.org/10.1007/s10462-022-10144-1.

[3] M. A, N.F. Ashi Mohammed Matuq, Siddiqui, Pre-trained word embeddings for arabic aspect-based sentiment analysis of airline tweets, in: M. F, S. K, A.A. T. Hassanien Aboul Ella, Tolba (Eds.), Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018, Springer International Publishing, Cham, 2019, pp. 241–251. Ed.

[4] Z.A. Abutiheen, E.A. Mohammed, M.H. Hussein, Behavior analysis in arabic social media, Int J Speech Technol 25 (3) (Sep. 2022) 659–666, https://doi.org/10.1007/s10772-021-09856-6.

[5] H.H. Hussein, A. Lakizadeh, IRAQIDSAD: a dataset for benchmarking sentiment analysis tasks on Iraqi dialect based texts, Int. J.Adv. Soft Comp. Appl. 16 (3) (Nov. 2024) 79–106, https://doi.org/10.15849/IJASCA.241130.06.

[6] B. Sabbar, N. Yousir, and L.A. Habeeb, 'Sentiment analysis for Iraqis dialect in social MEDIA using machine learning algorithms', 2018. [Online]. Available: https://ijict.edu.iq.

[7] N.T. Mohammed, E.A. Mohammed, H.H. Hussein, Evaluating various classifiers for Iraqi dialectic sentiment analysis. Lecture Notes in Networks and Systems, Springer Science and Business Media Deutschland GmbH, 2023, pp. 71–78, https://doi.org/10.1007/978-981-19-1412-6_6.

[8] A. Ali, J. Askar, N. Nur, and A. Sjarif, 'Annotated corpus of Mesopotamian-iraqi dialect for sentiment analysis in social Media', 2021. [Online]. Available: www.ijacsa.thesai.org.

[9] A. Alnawas, N. Arici, Sentiment analysis of Iraqi arabic dialect on Facebook based on distributed representations of documents, ACM Trans. Asian Low-Res. Language Inf. Process. 18 (3) (2019), https://doi.org/10.1145/3278605. Jan.

[10] M. Al-Jawad, H. Alharbi, A.F. Almukhtar, A.A. Alnawas, Constructing twitter corpus of Iraqi arabic dialect (CIAD) for sentiment analysis, Scien. Tech. J. Inform. Technol. Mech. Optics 22 (2) (Mar. 2022) 308–316, https://doi.org/10.17586/2226-1494-2022-22-2-308-316.

[11] N.D. Zaki, N.Y. Hashim, Y.M. Mohialden, M.A. Mohammed, T. Sutikno, A.H. Ali, A real-time big data sentiment analysis for iraqi tweets using spark streaming, Bull. Elect. Eng Inform. 9 (4) (Aug. 2020) 1411–1419, https://doi.org/10.11591/eei.v9i4.1897.

[12] A.R. Alfarhany, N.A.Z. Abdullah, Iraqi sentiment and emotion analysis using deep learning, J. Eng. 29 (09) (Sep. 2023) 150–165, https://doi.org/10.31026/j.eng.2023.09.11.

[13] W. Antoun, F. Baly, H.M. Hajj, AraBERT: transformer-based model for arabic language understanding, ArXiv abs/2003.00104 (2020) [Online]. Available, https://api.semanticscholar.org/CorpusID:211678011.

[14] H. Chouikhi, H. Chniter, F. Jarray, Arabic sentiment analysis using BERT model, in: International Conference on Computational Collective Intelligence, 2021 [Online]. Available, https://api.semanticscholar.org/CorpusID:238424899.

[15] R.A. Alsuhemi, S.M. Zarbah, Machine learning and AraBERT models for Arabic online reviews sentiment analysis, Romanian J. Informat Technol. Automatic Control (2022) 1–14.

[16] H. El-Moubtahij, H. Abdelali, E.B. Tazi, AraBERT transformer model for Arabic comments and reviews analysis, IAES International Journal of Artificial Intelligence (IJ-AI) (2022) [Online]. Available, https://api.semanticscholar.org/CorpusID:246222239.

[17] A. Humeau-Heurtier, Texture feature extraction methods: a survey, IEEE Access 7 (2019) 8975–9000 [Online]. Available, https://api.semanticscholar.org/CorpusID:59232541.

[18] A. Lakizadeh, Z. Zinaty, A novel hierarchical attention-based method for aspect-level sentiment classification, JAIDM (2021).

[19] S. Qaiser, R. Ali, Text mining: use of TF-IDF to examine the relevance of words to documents, Int J Comput Appl (2018) [Online]. Available, https://api.semanticscholar.org/CorpusID:53702508.

[20] G. Biau, E. Scornet, A random forest guided tour, TEST 25 (2015) 197–227 [Online]. Available, https://api.semanticscholar.org/CorpusID:14518730.

[21] Y. Song, Y. Lu, Decision tree methods: applications for classification and prediction, Shanghai Arch Psychiatry 27 (2015) 130–135 [Online]. Available, https://api.semanticscholar.org/CorpusID:18242585.

[22] H. Zhang, R. Zhou, The analysis and optimization of decision tree based on ID3 algorithm, in: 2017 9th International Conference on Modelling, Identification and Control (ICMIC), 2017, pp. 924–928, https://doi.org/10.1109/ICMIC.2017.8321588.

[23] S. Suthaharan, 'Machine learning models and algorithms for big data classification', 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:63539392.

[24] A. Patle, D.S. Chouhan, SVM kernel functions for classification, in: 2013 International Conference on Advances in Technology and Engineering (ICATE), 2013, pp. 1–9 [Online]. Available, https://api.semanticscholar.org/CorpusID:33346614.

[25] G.I. Webb, K. E, R. Miikkulainen, Naïve Bayes, Encycl. mach. learn. 15 (1) (2010) 713–714.

[26] H. Zhang, D. Li, Naïve Bayes text classifier, in: 2007 IEEE International Conference on Granular Computing (GRC2007), 2007, p. 708 [Online]. Available, https://api.semanticscholar.org/CorpusID:15601725.

[27] P. Cunningham, S.J. Delany, 'k-Nearest neighbour classifiers - A tutorial, ACM Computing Surveys (CSUR) 54 (2020) 1–25 [Online]. Available, https://api.semanticscholar.org/CorpusID:216641892.

[28] J.C. Stoltzfus, Logistic regression: a brief primer, Acad Emerg Med 18 (10) (2011) 1099. -104[Online]. Available, https://api.semanticscholar.org/CorpusID:33452324.

[29] D.G. Kleinbaum, K. Dietz, M. Gail, et al., Logistic Regression, Springer-Verlag, New York, 2002, pp. 43–53.

[30] J. Cheng, L. Dong, M. Lapata, Long short-term memory-networks for machine reading, ArXiv abs/1601.06733 (2016) [Online]. Available, https://api.semanticscholar.org/CorpusID:6506243.

[31] R. Dey, F.M. Salem, Gate-variants of Gated Recurrent Unit (GRU) neural networks, in: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 2017, pp. 1597–1600 [Online]. Available, https://api.semanticscholar.org/CorpusID:8492900.

[32] D. Gamal, M. Alfonse, E.-S.M. El-Horbaty, A.-B.M. Salem, Implementation of machine learning algorithms in arabic sentiment analysis using N-gram features, Procedia Comput Sci (2019) [Online]. Available, https://api.semanticscholar.org/CorpusID:199015677.

[33] Y. Li, Z. Hao, and L. Hang, 'Survey of convolutional neural network', 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:116282841.