International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 10, No. 3, June, 2023, pp. 790-798
Journal homepage: http://iieta.org/journals/mmep

Enhancing Arabic Sentiment Analysis in E-Commerce Reviews on Social Media Through a Stacked Ensemble Deep Learning Approach

Nouri Hicham[1*], Sabri Karim[1], Nassera Habbat[2]

[1] Research Laboratory on New Economy and Development (LARNED), Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University of Casablanca, Casablanca 20300, Morocco
[2] RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca 20300, Morocco

Corresponding Author Email: nhicham191@gmail.com

https://doi.org/10.18280/mmep.100308

Received: 24 January 2023
Accepted: 15 March 2023

Keywords:
AraBERT, deep learning, sentiment analysis, stacked model

ABSTRACT

Sentiment analysis (SA) employs natural language processing techniques to extract opinions from textual data. Applying SA to the Arabic language presents numerous challenges, including ambiguity, the presence of multiple dialects, a need for additional resources, and morphological variation. The domain of Arabic SA has witnessed significant advancements with the application of deep learning (DL) approaches, such as convolutional neural networks (CNNs). The performance of single DL models has been further improved by hybrid models combining CNNs with bidirectional long short-term memory (Bi-LSTM) or bidirectional gated recurrent units (Bi-GRU). It is anticipated that the accuracy of these DL models can be enhanced through stacked deep learning ensembles. In this study, a stacked ensemble approach is proposed that accurately predicts Arabic sentiment by leveraging the predictive capabilities of CNN, Bi-GRU, Bi-LSTM, and hybrid DL models (CNN-Bi-GRU and CNN-Bi-LSTM). The proposed model's efficacy is evaluated using four extensive datasets: the HARD dataset, the BRAD dataset, the ARD dataset, and a real dataset composed of 71,583 Arabic reviews. Experimental results demonstrate the suitability of the proposed model for analyzing sentiments in Arabic texts. The method's first step involves feature extraction using the AraBERT model. Subsequently, five DL models are developed and trained, including CNN, Bi-GRU, Bi-LSTM, a hybrid CNN-Bi-GRU model, and a hybrid CNN-LSTM model. Finally, the outputs of the base classifiers are concatenated using the multilayer perceptron algorithm. Our approach achieves an improved accuracy of 0.9256 compared to basic and hybrid deep learning methods.

1. INTRODUCTION

Over the past decade, the popularity of social media platforms and applications, such as Twitter, LinkedIn, and Facebook, has surged significantly. Businesses and various organizations have recognized the value of information obtained through engaging with and learning about their customers on these platforms. However, assessing a user's overall satisfaction with a brand becomes challenging due to the sheer volume of posts, users, messages, comments, and other forms of communication [1]. Sentiment analysis (SA), a subfield of natural language processing (NLP), employs advanced machine learning and data mining techniques to evaluate attitudes, emotions, reactions, and sentiments across diverse domains, including service quality, market reach, pricing, and public support for governmental events and actions [2].

Arabic, characterized by various irregular forms, complex morpho-syntactic agreement rules, and a wide range of linguistic variations, poses unique challenges for creating generalized models without proper handling and processing of Arabic text. Moreover, Arabic Sentiment Analysis (ASA) has fewer available resources, such as sentiment lexicons and annotated sentiment corpora, compared to English Sentiment Analysis. Consequently, ASA has garnered significant attention in recent years [3].

Deep learning (DL) and machine learning (ML) approaches facilitate the automation of extracting meaningful insights and sentiments from vast amounts of text [4, 5]. Recently, convolutional neural networks (CNNs) and hybrid models, such as those combining CNNs with long short-term memory (LSTM) networks, have improved sentiment analysis performance [6]. CNNs extract valuable information from text data by utilizing deep layers, including pooling, convolutional, and fully connected layers. LSTMs, with their memory state capabilities, can effectively memorize crucial information from the text and comprehend sentence meanings as a whole. For instance, Al Omari et al. [7] proposed a CNN-LSTM hybrid model using a word2vec model to classify Arabic sentiments into two main categories, combining the strengths of both CNNs and LSTMs. Similarly, Yang et al. [8] suggested employing a hybrid LSTM-RNN model for Arabic sentiment analysis, built upon LSTM and RNN architectures.

Ensemble learning techniques have gained prominence in machine learning research in recent years, yet their application in opinion mining remains relatively limited, particularly regarding heterogeneous ensembles constructed using various base classifiers and datasets [9]. Ensemble classifiers merge decisions made by multiple classifiers into a single model, anticipating that the combined model will produce superior results compared to each base classifier [10]. Ensemble modeling has become an increasingly popular method for enhancing the overall performance of NLP models. However, this training-based ensemble tends to overfit in situations with limited data [11, 12].

There is an increase in variance between base classifiers in heterogeneous and homogeneous ensembles due to various approaches [13]. Heterogeneous ensembles are prone to many different forms of bias, and if the biases complement one another, aggregating these biased judgments can outperform homogenous ensembles. Bagging [14], boosting [15], and stacking [16] are the three primary categories of ensemble learning approaches that can be applied to a given dataset.

In this study, a stacked ensemble model for ASA is proposed, utilizing Multilayer Perceptron (MLP) and SVM as meta-learners. This model is based on the optimal combination of hybrid CNN-BiLSTM, CNN, Bi-LSTM, Bi-GRU, and CNN-BiGRU architectures. After the optimization process, the performance of the constructed model surpasses that of other models in comparison. Our key contributions can be summarized as follows:

Proposal of hybrid CNN-BiLSTM, CNN-BiGRU, CNN, Bi-LSTM, and Bi-GRU deep learning architectures.

Development of a stacked ensemble model that combines the pretrained hybrid CNN-BiLSTM, CNN, CNN-BiGRU, Bi-LSTM, and Bi-GRU deep learning models. The outputs of the five different deep learning base classifiers are combined using Multilayer Perceptron, which acts as a meta-learner.

Comparison of the stacking model's performance with various deep learning models using several word embedding models (MLP and SVM) to determine the superiority of the proposed ensemble model in terms of accuracy, precision, recall, F1-score, and specificity. The suggested ensemble stacking model achieved significantly better results than existing deep learning models.

The remainder of this paper is organized as follows: Section 2 discusses the relevant sentiment analysis models for Arabic; Section 3 presents an overview of the proposed model; Section 4 provides a detailed analysis of the experimental results; and the final section concludes the paper.

2. RELATED WORKS

The majority of existing techniques for Arabic Sentiment Analysis (ASA) primarily focus on traditional machine learning methods. For instance, Omari [17] applied Logistic Regression (LR) to classify customer reviews, achieving the best results using the Term Frequency-Inverse Document Frequency (TF-IDF) representation. Similar methodologies have been employed in numerous Arabic and dialect sentiment classification studies. In a related work, Hadwan [18] utilized various machine learning techniques to analyze the opinions of Saudi Arabian people on social media. Through their experimental studies, it was found that the K-Nearest Neighbor (KNN) algorithm outperformed Support Vector Machine (SVM), Decision Tree (DT), and Naive Bayes (NB), reaching an accuracy of 78.46%. In the study [19], machine learning algorithms such as NB, Rocchio classifier, and SVM were used to assess Subjectivity and Sentiment Analysis of customer reviews in Arabic. The results demonstrated that the proposed ensemble method, with LR as the meta-learner, achieved superior performance.

In contrast to traditional ML approaches, recent studies have explored the potential of deep learning (DL) techniques for Sentiment Analysis [20, 21], as they offer increased robustness and adaptability through automatic feature extraction. Such techniques include Recurrent Neural Networks (RNN) [22] and Convolutional Neural Networks (CNN) [23]. However, the application of deep neural network (DNN) techniques to Arabic dialect Sentiment Analysis remains limited compared to their use in other domains such as chatbots, remote sensing, recommendation systems, and load monitoring.

In the study [24], both CNN and Long Short-Term Memory (LSTM) networks were employed, and an embedding matrix for words was generated using word2vec. This model achieved optimal performance on the SemEval 2017 and ASTD datasets. A CNN-LSTM hybrid model incorporating word2vec extracted features for binary categorization of Arabic opinions was proposed in the study [8]. Various ASA datasets, including Main-AHS, ASTD, and Ar-Twitter, were utilized, with the CNN-LSTM hybrid model attaining the highest accuracy of 79.07%. In the study [25], binary sentiment analysis was conducted using CNN on nine different datasets, including ASTD and LABR, which consist of tweets and reviews. Two types of word2vec, Skip-Gram and Continuous Bag-of-Words (CBOW), were used to create the word embedding matrix. Furthermore, CNN was applied to both balanced and unbalanced datasets. In the study [9], a hybrid LSTM-RNN model based on LSTM and RNN was proposed for Arabic sentiment analysis and the impact of using various pre-trained word embeddings with deep learning on the results was investigated. The AraSenTi-Tweet dataset was used to evaluate the model.

Ensemble models have the potential to enhance the inference power of individual models. Furthermore, the performance of hybrid techniques might be improved by employing hybrid models as base classifiers within an ensemble [10]. Ensemble models have been applied in various domains, often yielding superior results compared to base models [26]. In the context of ASA, ensemble modeling has also been utilized. For example, Saleh et al. [10] developed an ensemble model using voting to optimize Arabic sentiment analysis. The ASTD dataset was employed along with a CNN-LSTM model and an optimization strategy to select the most effective LSTM and CNN, based on the highest F1-score. The results indicated that the ensemble model exhibited the highest accuracy and F1-score, outperforming the individual models.

In conclusion, the body of related work demonstrates that both traditional machine learning and deep learning techniques have been explored for Arabic Sentiment Analysis. However, the application of deep learning techniques remains limited, and ensemble models have only recently been employed in this domain. The literature reviewed here provides a solid foundation upon which the proposed stacked ensemble model for ASA can be built and evaluated.

3. METHODOLOGY

In this study, we proposed a novel ASA approach founded on the stacked ensemble learning concept. The deep learning classifiers used in the proposed technique include hybrid CNN-BiLSTM, CNN-BiGRU, and CNN architectures. The suggested method then combines the results of these classifiers with meta-classifiers. By employing this method of stacked ensemble learning, we can improve overall performance while simultaneously capitalizing on each model's structural and functional benefits. In the following paragraphs, we will go through the hybrid models, the word embedding model, and the base and meta classifiers and go into additional detail regarding each.

3.1 Word embedding

Word embedding is a method that can be used to represent words in a numerical vector space. It helps us to retain the word's meaning in a way that is efficient from a computational standpoint. Word embedding is a method that maps individual words to a vector space, where each vector stands for a different phrase or word. This vector space is then utilized to describe the relations between the words and phrases, which assists us to grasp the contexts in which they are employed.

3.1.1 GloVe

The purpose of GloVe [27] is to investigate the global presentation of the whole corpus and to integrate the meaning of the terms into this. Word frequency and co-occurrence are the primary metrics utilized while determining the value of the real-valued vectors associated with specific words. The gloVe is an unsupervised method, meaning no human is involved in adding meaning to the collection of words. The utilization of the frequency of particular words and the frequency of the words immediately adjacent to each word serves as the foundation for the computation. The first item that must be completed in GloVe is compiling a list of the most frequently used terms as the context. The second step is constructing a co-occurrence matrix Z by reviewing the corpus and scanning the terms. Let us use n to represent the index of frequently occurring words and m to represent the remaining words in the corpus. P_{nm} is the probability that the given word m will appear in the same context as the given word n.

$$P_{nm} = \mathrm{Z}_{nm} / Z_n \tag{1}$$

Using *n*, *m*, and a third context word, *k*, we may determine the probability of co-occurrence $R_{(n,\, m,\, k)}$ as follows:

$$R_{(n,m,k)} = \frac{\mathrm{P}_{nm}}{\mathrm{P}_{mk}} \tag{2}$$

To conclude, the loss function *L* can be computed as follows:

$$L = \sum_{n,m=1}^{v} f(Z_{nm})(W_n^T k + B_n + B_m - \log Z_{nm})^2 \tag{3}$$

where, *f* is the function for giving weights, the training aims to reduce the least squares error as much as possible. After GloVe has been trained, each word is given a real-valued vector.

3.1.2 FastText

The fastText open-source project developed by Facebook Research is a popular NLP method that classifies and quickly represents text. The primary focus of the fastText embedding is not on learning word representations but on examining the underlying structure of words. This works incredibly well in languages with many morphemes since it frees students to memorize their representations of words with many morphemes [28]. By utilizing Skip-Gram, the probability of the context, which is denoted by a word t, is parameterized by word vectors through the utilization of a scoring function *S*:

$$S(w_t, w_c) = \mathrm{U}_{wt}^T \, \mathrm{V} \, wc \tag{4}$$

With *U* and *V* selected from the input and output matrix embeddings. fastText's scoring function is as follows:

$$S(w_t, w_c) = \sum_{g \in Gwt} \Lambda_g^T \, \mathrm{V} \, wc \tag{5}$$

Gwt is shorthand for the collection of n-grams found in the word w_t, and Λ_g is the vector representation of the gth n-gram. V_{wc} is the notation that indicates the vector associated with the context word w_c.

3.1.3 AraBERT

AraBERT is an Arabic pre-training BERT transformer model that uses word embedding to express semantics in context. It was trained on datasets from Arabic news websites: 1 billion tokens in 3.5 million articles from OSLAN Corpus and 1.5 billion words in 5 million articles from 10 primary news sources from 8 countries. Figure 1 shows the best-as-a-service technology, which activates layers without fine-tuning AraBERT settings [29]. It estimates the second-to-last concealed token pool's average. The output representation is fed into the categorization models we will discuss next.

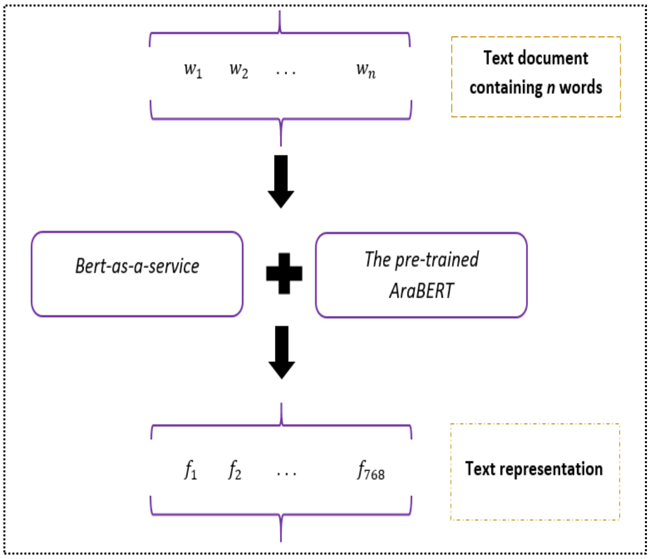

Figure 1. AraBERT feature extraction

3.2 Deep learning techniques

We presented in this part the deep learning models: a standard convolutional neural network (CNN), Bi-LSTM, Bi-GRU, a hybrid of a CNN and a Bi-LSTM network, and a hybrid of a CNN and a Bi-GRU network. The subsequent text provides a comprehensive analysis of each model.

3.2.1 Convolutional neural network (CNN)

CNN is primarily used for computer vision, but it has recently been expanded to NLP problems and achieved outstanding results in Arabic. CNNs were introduced to our opinion target extraction model to get character-level characteristics like suffixes and prefixes [30]. CNN-trained character vectors. By stacking the searched character vectors, matrix C is formed. Then, to get the most out of the pooling of character-level data, various convolution filters of variable widths are placed between matrix C and a large number of filter matrices. To prevent overfitting, we took advantage of the dropout layer in CNN before introducing character embedding.

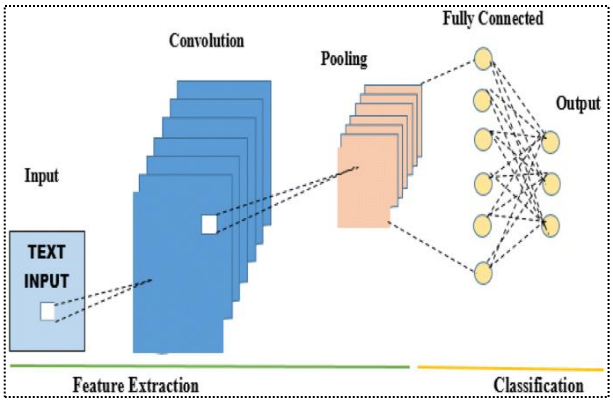

Figure 2. CNN sentiment analysis architecture

Figure 2 illustrates this point. The process of extracting features from text is carried out in an automated fashion by a convolution layer, which makes use of a variety of filters. In order to minimize the size of the feature maps, a pooling layer is utilized during the feature reduction process. Finally, a fully connected structure between the artificial neurons symbolizes the features of the text and the corresponding target tags to calculate the forecast class probability.

3.2.2 Bidirectional GRU (Bi-GRU)

The fact that GRU networks cannot exploit the current or future context results in the loss of information. Process data in both directions with the help of bidirectional GRU (Bi-GRU), which many researchers have employed. The hidden levels' information is brought up to the output layer. Two GRUs are brought together to form a bidirectional GRU network. While the input sequence of one network is shown in standard time order, the sequence of another network is presented in the opposite direction. At each stage, the outputs of both networks are combined [31]. This structure provides context, as shown in Figure 3.

3.2.3 Bidirectional LSTM (Bi-LSTM)

The BiLSTM [32] network is a subtype of the more general LSTM network. It employs a bidirectional network, which indicates that the inputs will run in two directions, as shown in Figure 4, one from the future to the past and the other from the past to the future. This allows the network to learn in both forward and backward directions. Consequently, it can keep the knowledge relevant to both the past and the future at any one point in time by utilizing the two hidden states simultaneously while simultaneously storing information in memory relevant to the future.

3.2.4 Hybrid networks

The combination of CNN, Bi-LSTM, and Bi-GRU models each has its distinct benefits and unique architecture:

* Utilized Bi-LSTM and Bi-GRU models to save data in the memory from the future while concurrently employing both hidden states to save data from the past and future.

*CNN is well-known for its ability to extract the most significant number of features feasible from the input data using the max-pooling layer.

The input layer of the Bi-LSTM and Bi-GRU model typically takes vectors as input from separate contextualized word embedding. The output of the Bi-LSTM and Bi-GRU model subsequently becomes the input of the CNN model. The goal of integrating these two architectures is to create a hybrid model that uses the benefits offered by CNN and those offered by the Bi-LSTM and Bi-GRU models. Following each filter, a max pooling layer is performed to decrease the amount of data and maintain its current state. After concatenation and linking, the outputs of these final layers are combined into a single two-dimensional softmax output. Ultimately, we apply the sigmoid activation function to datasets to obtain the predicted outcome and categorize datasets according to polarity. This function assigns binary labels to datasets. This action is taken to accomplish the goals that have been set.

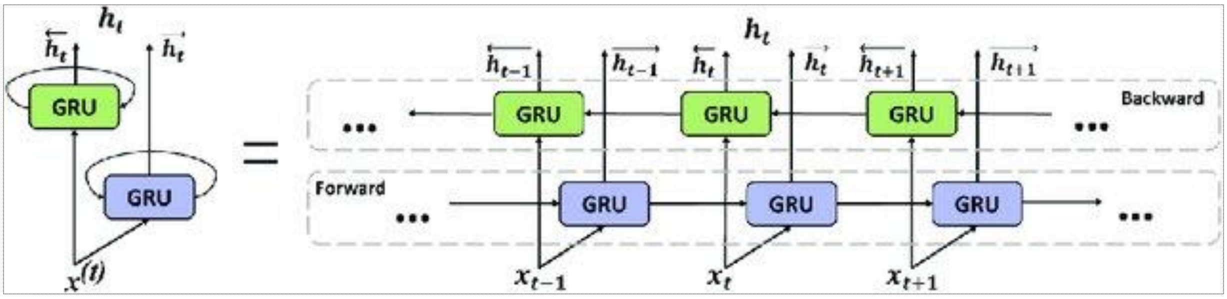

Figure 3. Bi-GRU structure

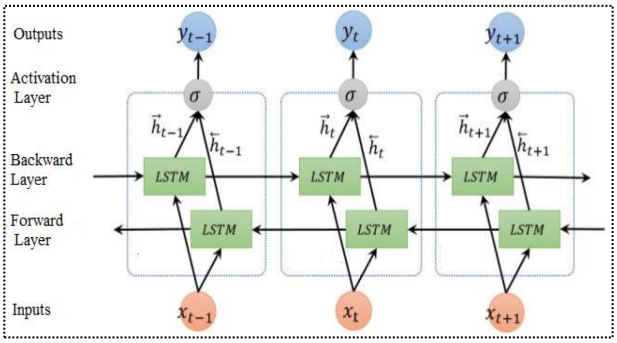

Figure 4. Bi-LSTM structure

4. EXPERIMENTS AND RESULTS

In the next section, we will first discuss the evaluation metric we employed to analyze our model's performance and then move on to the findings we obtained. When that is done, we will discuss our final thoughts.

4.1 Dataset description

The gathering of textual data is the first step in sentiment analysis. For the aim of assessing the effectiveness of our model in this research, the following datasets were utilized like shown in Table 1:

- The Hotel Arabic-Reviews Dataset (HARD) [33] includes 490587 hotel comments obtained from Booking's website during June and July 2016. The reviews are written in Modern Standard Arabic (MSA) and dialectal Arabic (DA), and the number of stars awarded by reviewers ranges from one to ten. Our research was based on the comprehensive HARD data set, which contained both positive and negative comments. The total number of reviews in the dataset is 93700, with an equal number of positive (46850) and negative (46850) types.

- The Books Reviews in Arabic Dataset (BRAD) [34] includes 510,600 book reviews. The reviews were collected from the GoodReads.com website during June and July 2016. MSA and DA constitute the majority of the research. We utilized the BRAD dataset, which contains equal amounts of positive and negative feedback. The only reviews shown are positive (four and five stars) and negative (one and two stars). The dataset consists of slightly more than 156 thousand reviews.

- The Arabic Reviews Dataset (ARD) [35] includes feedback from over 100,000 customers on hotels, films, books, and other things, as well as on chosen airlines. It falls into two major types (negative and positive).

Table 1. Description of the datasets

E-Commerce Reviews	Arabic datasets		
	HARD	BRAD	ARD
Negative	46 850	255 300	50 000
Positive	46 850	255 300	50 000
Total	93 700	510 600	100 000

4.2 Performance measures

The effectiveness of the suggested enhancement to our model's performance is evaluated using several different metrics. Metric selection can influence how the effectiveness and efficiency of models are tracked and compared. A summary of the five measures we used to rate the quality of our research is provided in Table 2.

Table 2. Description of sentiment analysis evaluation metrics

Performance evaluation	Summary	Equation
Accuracy	Accuracy is the ratio of instances properly predicted to the total number of instances.	$\frac{Tp+Tn}{Tp+Tn+Fp+Fn}$
Precision	Precision is the proportion of positively predicted samples relative to the total number of samples.	$\frac{Tp}{Tp+Fp}$
F1-score	The F1-score evaluates a player's precision and memory by computing the harmonic mean of their scores.	$\frac{2*(\text{Precision.Recall})}{(\text{Precision+Recall})}$
Recall	The percent of positive samples that were correctly expected is referred to as the recall.	$\frac{Tp}{Tp+Fn}$
Specificity	Specificity is the exact opposite of recall.	$\frac{Tn}{Tn+Fp}$

4.3 Experimental parameters

In our experiments, we utilized the following tested parameters as shown in Table 3.

Table 3. Experimental parameters

Setting parameter checked	Score ranges	Best value
Batch size	100. 80. 64. 32. 16. 8	64
Epoch	40. 30. 20. 15	15
Optimizer	RMSprop, SGD, Adadelta, Adagrad, Adamax, Adam, Adadelta, Nadam,	Adam
Dropout	0,6. 0,5. 0,4. 0,2. 0,0	0,4
Activation function	softsign, softmax, relu, softplus, linear, tanh	softplus

4.4 Experimental results

In this part, we discuss the findings of our studies that compared the impact of several word embeddings on the categorization of Arabic sentiment analysis performed with CNN, BiLSTM, BiGRU, CNN-BiLSTM, and CNN-BiGRU. Accuracy, Specificity, Precision, F1-score, and Recall were the measures we used to evaluate performance.

Tables 4, 5, and 6 display the obtained results for the datasets containing Arabic text employed. The findings show that the suggested model, which makes use of an ensemble learning technique with an MLP as a meta-classifier, outperforms the base classifiers with the highest accuracy of 92.03%, 92.58%, and 92.27% on the HARD, BRAD, and ARD datasets, respectively, when using an MLP word embedding model. This is demonstrated by the model being superior to the base classifiers. Compared to the CNN that used AraBERT, the proposed model has demonstrated a 13.4% increase in accuracy concerning the HARD dataset.

Analyzing the performance measures of the trained models with the HARD dataset reveals that our stacked model with MLP meta-learner and AraBERT word embedding had the most remarkable accuracy of 92.03%, followed by the SVM word embedding model with 91.86%, as shown in Table 4. The Bi-LSTM model with Glove word embedding registered the lowest accuracy at 73.15%.

As demonstrated in Table 5, our stacked model with MLP meta-learner and AraBERT word embedding had the highest accuracy of 92.58%, followed by the SVM word embedding model with 92.46%. At 70.28 per cent, the CNN model with Glove word embedding had the lowest accuracy.

As shown in Table 6, our stacked model consisting of an MLP meta-learner and an AraBERT word embedding achieved the most remarkable accuracy of 92.27%, followed by the SVM word embedding model, which achieved 92.15%. The CNN model that used the Glove word embedding had the lowest accuracy, reaching 69.97%.

Concerning the other evaluation metrics, which are precision, F1-score, Recall, and specificity, our model, which stacked hybrids models of deep learning; CNN-BiLSTM, CNN-BiGRU, and simple models like Bi-GRU and Bi-LSTM with the word embedding MLP, presents good results in all these metrics which are 94.12%, 72.22%, 88.02%, and 90.11% in ARD dataset, and 94.43%, 72.53%, 88.33%, 90.42% respectively in BRAD dataset, and for HARD dataset presents a performance that rises 93.88%, 71.98%, 87.78%, 89.87% in precision, Recall, F1-score, and specificity respectively.

4.5 Use case

In this part, we will implement the proposed (AraBERT + stacked model using hybrid models and MLP classifier) in a real dataset. This dataset comprises 71 583 Arabic reviews crawled from the Facebook page between 01 April 2022 and 01 September 2022 concerning Lesieur Morocco. Lesieur Morocco was initially established in 1941 by the Lesieur France Group, and in the 1980s, it was acquired by the National Group ONA.

The company manufactures and distributes a diverse selection of cooking oils, olive oils, margarine, and soaps, including brands such as Cristal, Lesieur, Al Horra, Huilor, Magdor, Mabrouka, Famila, Taous, Ledda, El Menjel, and El Kef. These brands have earned the support and continued loyalty of several generations of customers [36].

We collected reviews from the Lesieur Morocco Facebook page using web scraping techniques (the BeautifulSoup and Request Python libraries) and then stored them in the MongoDB database. Following that, we used MLP to extract features, and we carried out a sentiment analysis using a stacked ensemble hybrid deep learning model to get positive and negative reviews. We used the ProdLDA topic model to extract topics based on positive and negative classes for each topic.

The findings of the sentiment analysis are displayed in Figure 5, and a word cloud of the subjects generated by positive and negative sentiment is displayed in Figures 6 and 7.

Linked to the product's high prices, notably oil, a mainstay in Moroccan families, as seen in Figure 6. WordCloud of negative sentiments about the product's high pricing, On the other hand, the word cloud of positive attitudes that are displayed in Figure 7 demonstrates that the majority of consumers are pleased, particularly with the age of the brand, the quality of the products, and the variety of products available.

Table 4. Performance evaluation of applied models in HARD dataset

		HARD				
Deep learning techniques	**Word embeddings**	**Accuracy**	**Precision**	**Recall**	**F1-measure**	**Specificity**
CNN	**Glove**	0.6973	0.7488	0.4321	0.6884	0.8297
	FastText	0.7489	0.8075	0.5077	0.7011	0.8472
	AraBERT	0.7863	0.8409	0.6414	0.7278	0.8187
BiLSTM	**Glove**	0.7315	0.7604	0.4322	0.6689	0.8521
	FastText	0.8115	0.8697	0.5008	0.7223	0.7956
	AraBERT	0.8331	0.888	0.5878	0.7548	0.8338
BiGRU	**Glove**	0.7723	0.8185	0.6386	0.7704	0.8297
	FastText	0.8334	0.836	0.6946	0.6659	0.8472
	AraBERT	0.8863	0.9144	0.5977	0.6019	0.8193
CNN-BiLSTM	**Glove**	0.7331	0.7736	0.5869	0.7992	0.8801
	FastText	0.8677	0.9126	0.5317	0.7277	0.7658
	AraBERT	0.8985	0.9311	0.6269	0.8519	0.8187
CNN-BiGRU	**Glove**	0.8901	0.7488	0.4321	0.6884	0.7798
	FastText	0.7489	0.8075	0.5077	0.6011	0.8344
	AraBERT	0.7863	0.8409	0.6414	0.7278	0.7798
Stacked model with MLP	**Glove**	0.7831	0.8361	0.7094	0.7289	0.8197
	FastText	0.8414	0.8875	0.5669	0.7321	0.6473
	AraBERT	**0.9203**	**0.9388**	**0.7198**	**0.8778**	**0.8987**
Stacked model with SVM	**Glove**	0.7814	0.8344	0.7077	0.7272	0.818
	FastText	0.8397	0.8858	0.5652	0.7304	0.6456
	AraBERT	0.9186	0.9371	0.7181	0.8761	0.897

Table 5. Performance evaluation of applied models in BRAD dataset

		BRAD				
Deep learning techniques	**Word embeddings**	**Accuracy**	**Precision**	**Recall**	**F1-measure**	**Specificity**
CNN	**Glove**	0.7028	0.7543	0.4376	0.6939	0.8352
	FastText	0.7544	0.8139	0.5132	0.7066	0.8527
	AraBERT	0.7918	0.8464	0.6469	0.7333	0.8242
BiLSTM	**Glove**	0.7371	0.7659	0.4377	0.6744	0.8576
	FastText	0.8172	0.8752	0.5063	0.7278	0.8011
	AraBERT	0.8386	0.8935	0.5933	0.7603	0.8393
BiGRU	**Glove**	0.7778	0.8247	0.6441	0.7759	0.8352
	FastText	0.8389	0.8415	0.7001	0.6714	0.8527
	AraBERT	0.8918	0.9199	0.6032	0.6074	0.8248
CNN-BiLSTM	**Glove**	0.7386	0.7791	0.5924	0.8047	0.8856
	FastText	0.8732	0.9181	0.5372	0.7332	0.7713
	AraBERT	0.9045	0.9366	0.6324	0.8574	0.8242
CNN-BiGRU	**Glove**	0.8956	0.7543	0.4376	0.6939	0.7853
	FastText	0.7544	0.8135	0.5132	0.6066	0.8399
	AraBERT	0.7918	0.8464	0.6469	0.7333	0.7853
Stacked model with MLP	**Glove**	0.7886	0.8416	0.7149	0.7344	0.8252
	FastText	0.8469	0.8937	0.5724	0.7376	0.6528
	AraBERT	**0.9258**	**0.9443**	**0.7253**	**0.8833**	**0.9042**
Stacked model with SVM	**Glove**	0.7874	0.8404	0.7137	0.7332	0.824
	FastText	0.8457	0.8925	0.5712	0.7364	0.6516
	AraBERT	0.9246	0.9431	0.7241	0.8821	0.903

Table 6. Performance evaluation of applied models in ARD dataset

		ARD				
Deep learning techniques	**Word embeddings**	**Accuracy**	**Precision**	**Recall**	**F1-measure**	**Specificity**
CNN	**Glove**	0.6997	0.7512	0.4345	0.6908	0.8321
	FastText	0.7513	0.8108	0.5101	0.7035	0.8496
	AraBERT	0.7887	0.8433	0.6438	0.7302	0.8211
BiLSTM	**Glove**	0.7347	0.7628	0.4346	0.6713	0.8545
	FastText	0.8141	0.8721	0.5032	0.7247	0.798
	AraBERT	0.8355	0.8904	0.5902	0.7572	0.8362
BiGRU	**Glove**	0.7747	0.8216	0.6417	0.7728	0.8321
	FastText	0.8358	0.8384	0.6973	0.6683	0.8496
	AraBERT	0.8887	0.9168	0.6001	0.6043	0.8217
CNN-BiLSTM	**Glove**	0.7355	0.7763	0.5893	0.8016	0.8825
	FastText	0.8701	0.9151	0.5341	0.7301	0.7682
	AraBERT	0.9014	0.9335	0.6293	0.8543	0.8211
CNN-BiGRU	**Glove**	0.8925	0.7512	0.4345	0.6908	0.7822
	FastText	0.7513	0.8104	0.5101	0.6035	0.8368
	AraBERT	0.7887	0.8433	0.6438	0.7302	0.7822
Stacked model with MLP	**Glove**	0.7855	0.8385	0.7118	0.7313	0.8221
	FastText	0.8438	0.8906	0.5693	0.7345	0.6497
	AraBERT	**0.9227**	**0.9412**	**0.7222**	**0.8802**	**0.9011**
Stacked model with SVM	**Glove**	0.7843	0.8373	0.7106	0.7301	0.8209
	FastText	0.8426	0.8894	0.5681	0.7333	0.6485
	AraBERT	0.9215	0.94	0.721	0.879	0.8999

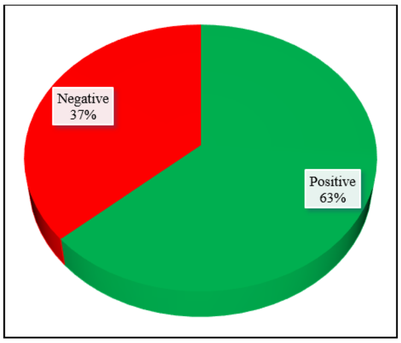

Figure 5. Reviews sentiment analysis of Lesieur Morocco

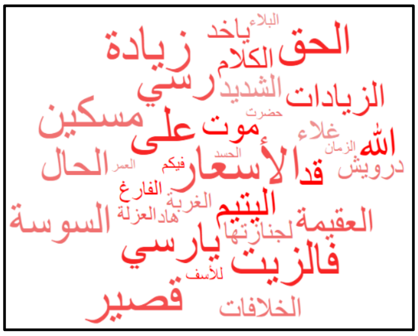

Figure 6. Wordcloud negative sentiments

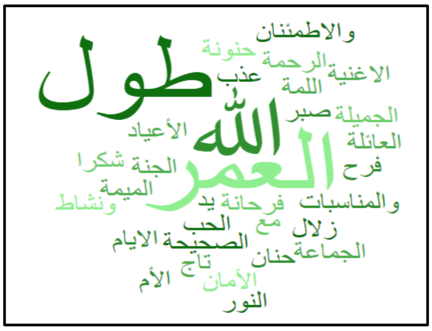

Figure 7. Wordcloud positive sentiments

5. CONCLUSION AND FUTURE DIRECTIONS

The challenge of analyzing the feelings conveyed in Arabic writing is investigated in this study. The effectiveness of the Arabic sentiment analysis system was evaluated concerning ensemble stacking models based on CNN, BiGRU, BiLSTM, hybrid CNN-BiGRU model, and hybrid CNN-BiLSTM model, respectively.

The proposed model's efficiency is measured by employing three enormous datasets: the HARD, BRAD, and ARD. The experimental results demonstrated that the suggested model is suitable for analyzing the sentiments contained in Arabic language texts. The proposed method's first step is extracting features with the Arabert model. Next, we develop and train five deep learning models, including CNN, BiGRU, BiLSTM, a hybrid CNN-BiGRU model, and a hybrid CNN-LSTM model. After that, the outputs of the underlying classifiers are concatenated using a meta-classifier called a Multilayer Perceptron algorithm.

Our methodology was validated using an authentic Arabic review dataset. The recommended tactic beats the baseline models on the BRAD dataset when applied using Arabert, with the best accuracy of 0.9256. The study's outcomes shed light on the value of textual data to professionals in developing strategies, enhancing competitiveness, and managing income. In upcoming work, the Arabic word's meaning is vital, and it is anticipated that this would lead to improved performance. We will take into consideration this constraint in upcoming work.

REFERENCES

[1] Habbat, N., Anoun, H., Hassouni, L. (2022). Combination of GRU and CNN deep learning models for sentiment analysis on French customer reviews using XLNet model. IEEE Engineering Management Review, 51(1): 41-51. https://doi.org/10.1109/EMR.2022.3208818

[2] El-Affendi, M.A., Alrajhi, K., Hussain, A. (2021). A

novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain Arabic sentiment analysis. IEEE Access, 9: 7508-7518. https://doi.org/10.1109/ACCESS.2021.3049626

[3] Badaro, G., Baly, R., Hajj, H., El-Hajj, W., Shaban, K.B., Habash, N., Al-Sallab, A., Hamdi, A. (2019). A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3): 1-52. https://doi.org/10.1145/3295662

[4] Wankhade, M., Rao, A.C.S., Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7): 5731-5780. https://doi.org/10.1007/s10462-022-10144-1

[5] Hicham, N., Karim, S. (2022). Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering. International Journal of Advanced Computer Science and Applications, 13(10). https://doi.org/10.14569/IJACSA.2022.0131016

[6] Al-Hashedi, A., Al-Fuhaidi, B., Mohsen, A.M., Ali, Y., Gamal Al-Kaf, H.A., Al-Sorori, W., Maqtary, N. (2022). Ensemble classifiers for Arabic sentiment analysis of social network (twitter data) towards covid-19-related conspiracy theories. Applied Computational Intelligence and Soft Computing, 2022: 6614730. https://doi.org/10.1155/2022/6614730

[7] Al Omari, M., Al-Hajj, M., Sabra, A., Hammami, N. (2019). Hybrid CNNs-LSTM deep analyzer for Arabic opinion mining. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, pp. 364-368. https://doi.org/10.1109/SNAMS.2019.8931819

[8] Yang, L., Li, Y., Wang, J., Sherratt, R.S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access, 8: 23522-23530. https://doi.org/10.1109/ACCESS.2020.2969854

[9] Heikal, M., Torki, M., El-Makky, N. (2018). Sentiment analysis of Arabic tweets using deep learning. Procedia Computer Science, 142: 114-122. https://doi.org/10.1016/j.procs.2018.10.466

[10] Saleh, H., Mostafa, S., Alharbi, A., El-Sappagh, S., Alkhalifah, T. (2022). Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis. Sensors, 22(10): 3707. https://doi.org/10.3390/s22103707

[11] Galal Elsayed, H.A., Chaffar, S., Brahim Belhaouari, S., Raissouli, H. (2022). A two-level deep learning approach for emotion recognition in Arabic news headlines. International Journal of Computers and Applications, 44(7): 604-613. https://doi.org/10.1080/1206212X.2020.1851501

[12] Hicham, N., Karim, S., Habbat, N. (2022). An efficient approach for improving customer Sentiment Analysis in the Arabic language using an Ensemble machine learning technique. In 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet), Marrakech, Morocco, pp. 1-6. https://doi.org/10.1109/CommNet56067.2022.9993924

[13] Ardabili, S., Mosavi, A., Várkonyi-Kóczy, A.R. (2019). Advances in machine learning modeling reviewing hybrid and ensemble methods. Preprints.org 2019, 2019080203. https://doi.org/10.20944/preprints201908.0203.v1

[14] Sagi, O., Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4): e1249. https://doi.org/10.1002/widm.1249

[15] Freund, Y., Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1): 119-139. https://doi.org/10.1006/jcss.1997.1504

[16] Sarkar, K. (2020). A stacked ensemble approach to bengali sentiment analysis. In: Tiwary, U., Chaudhury, S. (eds) Intelligent Human Computer Interaction. IHCI 2019. Lecture Notes in Computer Science, vol. 11886. Springer, Cham. https://doi.org/10.1007/978-3-030-44689-5_10

[17] Omari, M.A. (2022). OCLAR: Logistic regression optimisation for Arabic customers' reviews. International Journal of Business Intelligence and Data Mining, 20(3): 251-273. https://doi.org/10.1504/IJBIDM.2022.122177

[18] Hadwan, M., Al-Hagery, M., Al-Sarem, M., Saeed, F. (2022). Arabic sentiment analysis of users' opinions of governmental mobile applications. Computers, Materials and Continua, 72(3): 4675-4689. https://doi.org/10.32604/cmc.2022.027311

[19] Omar, N., Albared, M., Al-Shabi, A.Q., Al-Moslmi, T. (2013). Ensemble of classification algorithms for subjectivity and sentiment analysis of Arabic customers' reviews. International Journal of Advancements in Computing Technology, 5(14): 77-85.

[20] Elnagar, A., Al-Debsi, R., Einea, O. (2020). Arabic text classification using deep learning models. Information Processing & Management, 57(1): 102121. https://doi.org/10.1016/j.ipm.2019.102121

[21] Farha, I.A., Magdy, W. (2021). A comparative study of effective approaches for Arabic sentiment analysis. Information Processing & Management, 58(2): 102438. https://doi.org/10.1016/j.ipm.2020.102438

[22] Williams, R.J., Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. Neural Computation, 1(2): 270-280. https://doi.org/10.1162/neco.1989.1.2.270

[23] Zhang, A., Lipton, Z.C., Li, M., Smola, A.J. (2021). Dive into deep learning. arXiv preprint arXiv:2106.11342. https://doi.org/10.48550/arXiv.2106.11342

[24] Farha, I.A., Magdy, W. (2019). Mazajak: An online Arabic sentiment analyser. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp. 192-198. http://dx.doi.org/10.18653/v1/W19-4621

[25] Dahou, A., Xiong, S., Zhou, J., Haddoud, M.H., Duan, P. (2016). Word embeddings and convolutional neural network for Arabic sentiment classification. In Proceedings of Coling 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2418-2427.

[26] Kang, M., Ahn, J., Lee, K. (2018). Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 94: 218-227. https://doi.org/10.1016/j.eswa.2017.07.019

[27] Pennington, J., Socher, R., Manning, C.D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

https://doi.org/10.3115/v1/D14-1162

[28] Setyanto, A., Laksito, A., Alarfaj, F., Alreshoodi, M., Oyong, I., Hayaty, M., Alomair, A., Almusallam, N., Kurniasari, L. (2022). Arabic language opinion mining based on Long Short-Term Memory (LSTM). Applied Sciences, 12(9): 4140. https://doi.org/10.3390/app12094140

[29] Habbat, N., Anoun, H., Hassouni, L. (2021). A novel hybrid network for Arabic sentiment analysis using fine-tuned Arabert model. International Journal on Electrical Engineering and Informatics, 13(4): 801-812. https://doi.org/10.15676/ijeei.2021.13.4.3

[30] Luo, L.X. (2019). Network text sentiment analysis method combining LDA text representation and GRU-CNN. Personal and Ubiquitous Computing, 23(3-4): 405-412. https://doi.org/10.1007/s00779-018-1183-9

[31] Abdelgwad, M.M., Soliman, T.H.A., Taloba, A.I., Farghaly, M.F. (2022). Arabic aspect based sentiment analysis using bidirectional GRU based models. Journal of King Saud University-Computer and Information Sciences, 34(9): 6652-6662. https://doi.org/10.1016/j.jksuci.2021.08.030

[32] Schuster, M., Paliwal, K.K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11): 2673-2681. https://doi.org/10.1109/78.650093

[33] Elnagar, A., Khalifa, Y.S., Einea, A. (2018). Hotel Arabic-reviews dataset construction for sentiment analysis applications. In: Shaalan, K., Hassanien, A., Tolba, F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham. https://doi.org/10.1007/978-3-319-67056-0_3

[34] Elnagar, A., Einea, O. (2016). Brad 1.0: Book reviews in Arabic dataset. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir, Morocco, pp. 1-8. https://doi.org/10.1109/AICCSA.2016.7945800

[35] Arabic 100k Reviews. https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews, accessed on June 29, 2022.

[36] Présentation Lesieur Cristal, Lesieur Cristal. https://www.lesieur-cristal.com/notre-groupe/presentation-lesieur-cristal/, accessed on June 30, 2022.