Date of publication SEP-30, 2024, date of current version MAY-31, 2024.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.3.XXXX

Towards Improving E-Commerce Customer Review Analysis for Arabic Language Opinion Mining

NOURI HICHAM[1], HABBAT NASSERA[2]
[1]Research Laboratory on New Economy and Development (LARNED),
Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University of Casablanca, Morocco
[2]Faculty of Science and Technology of Settat, Hassan First University, Settat, Morocco

Corresponding author: Nouri Hicham (e-mail: nhicham191@gmail.com).

ABSTRACT In recent years, the rapid development of Internet-related technologies has facilitated the widespread adoption of online purchasing as a convenient means of satisfying consumer needs. Conducting sentiment analysis on user reviews on e-commerce platforms can substantially improve customer satisfaction. In order to resolve this issue, we propose a novel model for sentiment analysis that employs hybrid deep learning ensembles, combining RNN and TreeLSTM with AraBERT as the word embedding. Our research concentrates on creating a hybrid deep-learning model to predict Arabic sentiment accurately. We employ deep learning models with various word embeddings, such as RNN, LSTM, BiLSTM, TreeLSTM, RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM. Multiple open-access datasets are used to evaluate the efficacy of the proposed model, including the BRAD dataset, the ARD dataset, and merged datasets containing 610,600 items. The experimental findings indicate that our proposed model is well-suited for evaluating the sentiments expressed in Arabic texts. Our strategy starts with extracting features using the Arabert model, followed by developing and training five hybrid deep-learning models. We attained a significant accuracy improvement of 0.9409 when comparing our method to traditional and hybrid deep learning techniques. This demonstrates that our proposed model precisely analyzes sentiment. These findings are important for enhancing the comprehension of emotions conveyed in Arabic text and have practical implications for various applications, especially e-commerce. By accurately assessing sentiment, businesses can better comprehend customer preferences and enhance consumer satisfaction by enhancing their offerings.

KEYWORDS hybrid deep learning; sentiment analysis; Arabic language; classification; AraBERT.

I. INTRODUCTION

Integrating e-commerce into our daily lives has significantly transformed our shopping habits and interactions with businesses. The increasing prevalence of Internet shopping has led to a notable impact of customer reviews on consumer purchasing choices. Customer evaluations are widely recognized as a helpful and informative resource that aids prospective buyers in assessing the caliber and dependability of various items or services. The analysis of customer reviews has become imperative for firms to comprehend consumer happiness and make well-informed business decisions [1].

The study of customer reviews in Arabic presents distinct obstacles arising from a range of linguistic and cultural issues. Arabic is a Semitic language renowned for its intricate and multifaceted structure, encompassing several dialects and diverse modes of communication. The Arabic script displays a range of linguistic phenomena, encompassing diacritical signs, complex morphology, and the omission of vowel symbols. The attributes above need help analyzing Arabic and extracting significant insights from customer reviews [2].

Opinion mining, also known as sentiment analysis, is a specialized area within the science of natural language processing (NLP) dedicated to extracting and examining subjective information, views, and sentiments conveyed through textual data [3]. The objective of opinion mining is to ascertain a given text's polarity (positive, negative, or neutral) and subsequently categorize it based on this determination. The field of opinion mining has received significant attention concerning the English language, but the exploration of Arabic opinion mining remains at a nascent level of development [4].

The conventional methodologies for sentiment analysis involve utilizing machine learning algorithms, such as Support Vector Machines (SVM) and Naive Bayes classifiers. Typically, these approaches necessitate the process of human

feature engineering, wherein pertinent linguistic elements are chosen to capture moods adequately. Nevertheless, the procedure above can prove to be arduous and time-consuming, thus impeding the efficacy of emotion analysis in intricate languages such as Arabic [5].

Natural Language Processing (NLP) has significantly transformed due to recent progress in deep learning techniques. These developments have profoundly impacted different NLP applications, such as sentiment analysis. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated encouraging outcomes in autonomously acquiring pertinent features from unprocessed textual input. These models can comprehend a given text's context, semantics, and syntax, thus diminishing the necessity for manual feature engineering [6].

This research article presents a novel hybrid deep-learning model designed for sentiment analysis of customer reviews in Arabic. Deep learning models, including RNN, LSTM, BiLSTM, TreeLSTM, RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM, are utilized in our study to include local and global contextual information in Arabic text through the use of diverse word embeddings.

The hybrid model under consideration comprises two primary constituents: the RNN and TreeLSTM components. The RNN component is designed to capture local aspects inside the text through convolutional processes. On the other hand, the TreeLSTM component is specifically designed to collect long-term dependencies and contextual information by employing recurrent connections. Through the integration of these two elements, our model is capable of efficiently acquiring and portraying the intricate sentiment patterns that are inherent in Arabic customer reviews.

To assess the efficacy of our proposed model, a comprehensive series of experiments were conducted on publicly accessible Arabic customer review datasets. These datasets encompassed the BRAD and ARD datasets and merged datasets with a total of 610,600 items. In this study, we conduct a comparative investigation of our model's performance in sentiment analysis, juxtaposing it against conventional machine learning approaches and contemporary state-of-the-art deep learning models. The findings indicate that our hybrid model performs better than the baselines, as it achieves higher accuracy and demonstrates improved sentiment classification capabilities.

The research paper presents a summary of the contributions made in this study. In this study, we provide a novel hybrid deep-learning model for sentiment analysis on Arabic customer reviews. Our proposed model leverages the advantages of Recurrent Neural Networks (RNNs) and TreeLSTM to achieve improved performance. Extensive experiments are conducted to assess the efficacy of the proposed model, demonstrating its superiority compared to conventional machine learning methods and alternative deep learning models. In this study, we offer a comprehensive examination of the difficulties and potential advantages of sentiment analysis in the context of Arabic customer evaluations. Additionally, we emphasize the significance of devising tailored methodologies specifically designed for the Arabic language.

In summary, the analysis of customer reviews is of utmost importance in comprehending client preferences and attitudes about items or services in electronic commerce. Nevertheless, studying customer feedback in Arabic poses distinctive obstacles due to its intricate language and cultural attributes. Our proposed hybrid deep learning model aims to effectively tackle these problems and enhance the sentiment analysis of Arabic customer evaluations. The experimental findings provide evidence of the efficacy of our model, highlighting its capacity to improve decision-making procedures and enhance customer happiness within the e-commerce sector.

II. RELATED WORKS

Sentiment Analysis (SA) is a field of study that entails extracting specific information from textual data [7]. It relates to several academic areas, including text mining, computational linguistics, and NLP – various terms, including subjective analysis, emotion extraction, and opinion mining, known as SA [8]. The difference between positive and negative ideas is analyzed in Opinion Mining using the text given by individuals, while emotion extraction tracks various emotions. SA examines people's attitudes, emotions, views, assessments, and sentiments towards services, issues, products, events, subjects, organizations, and individuals. SNS such as Facebook, Twitter, and LinkedIn are valuable sites for gathering data and performing SA [9].

SA is helpful in various disciplines, including corporate operations and politics. Commercial operations enable organizations to automatically collect consumer feedback on their services or products. In politics, SA can be used to infer the public's orientation and forecast their reaction to political actions, which can aid in political decision-making. SA can be performed at various levels, including topic, sentence, and document.

While SA is important, more research on SA in Arabic is needed due to the Arabic language's complex script, morphology, and ambiguity [4, 10]. Unsupervised and supervised learning methods are used in SA approaches, with the former utilizing Machine Learning (ML) techniques with feature engineering. Because manual feature extraction is time-consuming and labor-intensive, Deep Learning (DL) techniques have been proposed to improve SA in the English language [9, 11]. Several studies have proposed using a cutting-edge word embedding techniques and recurrent neural networks such as LSTM and Bi-direction Gated Recurrent Unit (Bi-GRU) to develop DL techniques for SA in Arabic. Instead of employing low-quality Arabic stop words or an artificial collection of stop word lists, Automated Sentimental Refinement (ASR) has been utilized to remove stop words. DL-based sentimental analysis algorithms have also been presented to forecast the divergence of sentiments and opinions [4, 9].

Various research studies have used ensemble learning to predict sentiment analysis. For example, Al-Hashedi et al [7] employed five distinct classifiers as an ensemble approach to predict sentiment analysis of Arabic tweets connected to COVID-19: Nave Bayes (NB), Stochastic Gradient Descent (SGD), Random Forest (RF), Logistic Regression (LR), and a voting classifier. The tweets were classified as excellent or negative, and the voting classifier outperformed the other classifiers.

Alharbi et al. [11] introduced the DeepASA model, which included an input layer, hidden layers, and two types of Deep Learning (DL) networks: Gated Recurrent Units (GRU) and Short-Term Long Memory (LSTM). The last layer employed a voting method to improve the model's prediction performance. The studies were carried out on six Arabic datasets: book reviews, hotel reviews, restaurant reviews, product reviews, Twitter, and the Arabic Sentiment Treebank Dataset (ASTD).

The DeepASA model performed admirably.

On the ASTD Arabic sentiment analysis dataset, Oussous et al. [12] used a voting algorithm on three classifiers: Support Vector Machines (SVM), NB, and Maximum Entropy. The results showed that the voting algorithm was very accurate.

To classify the sentiment of Arabic text, Al-Saqqa et al. [13] suggested an ensemble of four Machine Learning (ML) classifiers – K-Nearest Neighbours (KNN), SVM, NB, and a majority voting algorithm. Movie reviews, ArTwitter, and a large-scale Arabic sentiment analysis dataset (LABR) were employed. The experiments demonstrated that the classifier ensemble outperformed the individual classifiers.

On the Arabic sentiment dataset, Al-Azani et al. [14] compared the performance of various ensemble learning techniques such as bagging, boosting, voting, stacking, and RF. The stacking ensemble technique delivered excellent results.

Other authors have used ensemble learning techniques to analyze sentiment in non-Arabic languages. For example, Sitaula et al. [15] created the NepCOV19 Tweets Nepali Twitter sentiment dataset and presented a CNN ensemble approach to gathering multi-scale information for better categorization. The authors used a variety of feature selection strategies to develop each method, including fastText, domain-specific, domain-agnostic, and multiple CNN models. The authors suggested a multi-channel CNN (MCNN) model in [16] to classify positive, neutral, and negative attitudes in the NepCOV19Tweets dataset. The suggested MCNN model achieved accuracy using a hybrid feature extraction strategy for semantic and syntactic characteristics.

Previous work, however, used homogeneous ensemble learning or hybrid models rather than heterogeneous ensemble learning. In this work, we suggested a heterogeneous ensemble deep learning model improve Arabic sentiment analysis. To improve the model's performance for predicting Arabic sentiment analysis, the suggested model incorporated three different DL models, RNN, LSTM, and GRU, as well as three meta-learners, LR, RF, and SVM.

III. THE PROPOSED MODEL

The investigation resulted in the development of a novel strategy for the classification of feelings in Arabic. The model shown in the research can recognize and classify various emotions communicated via the use of the Arabic language. Consequently, we designed a hybrid deep learning model for ASA that uses the word embedding capabilities provided by both BERT, AraBERT and mBERT. The model is built with the help of the most potent combination of hybrid RNN-TreeLSTM. In addition, we propose a practical model that investigates pre-trained deep learning models, such as RNN, LSTM, BiLSTM, TreeLSTM, RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM. This model includes all of these models. The model utilizes word embedding methods such as FastText and Glove to aggregate the results obtained from the many different deep-learning databases.

A. WORD EMBEDDING

Converting words into numerical vectors is called word embedding and is utilized in natural language processing. The purpose of this method is to represent words in a space with many dimensions, where each dimension represents a different facet of the word's meaning. This method assists machines in gaining a more accurate understanding of the context in which words are used and their meaning. Several natural language processing applications extensively use the word embedding technique, such as sentiment analysis, machine translation, and speech recognition. It has been demonstrated to be an effective method for increasing the precision of NLP models by deciphering the meanings behind individual words' surfaces [14, 17].

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a Google-developed deep learning model for natural language processing applications. It employs a transformer architecture to encode the bidirectional context of words in a phrase. To discover the underlying language patterns, BERT is trained on a massive corpus of text data using unsupervised learning [18]. The BERT model can be fine-tuned for specific downstream NLP applications, including question answering, sentiment analysis, and language translation. BERT has produced cutting-edge results on various NLP benchmarks and is widely used in academic research and industrial applications. One of BERT's key features is its ability to recognize the context of words in a sentence, which aids in capturing natural language nuances. The development of BERT has opened up new avenues for language comprehension and resulted in considerable advances in various NLP tasks [18].

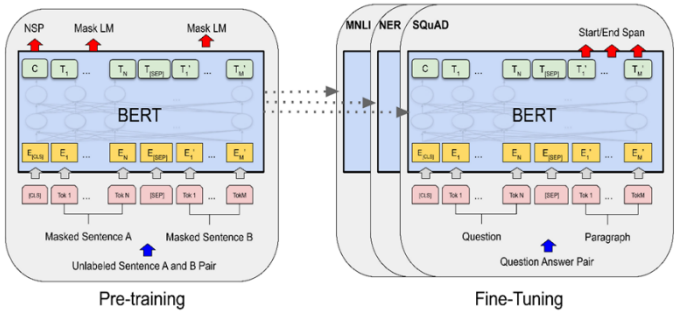

Figure 1. Bert structure

AraBERT

AraBERT [6] is a deep learning model that has already been pre-trained and is based on the BERT architecture. It was built explicitly for Arabic. It is educated on a sizable body of Arabic text data and then fine-tuned for various natural language processing (NLP) tasks, including sentiment analysis, named entity recognition, and machine translation. The development of AraBERT is significant because the Arabic language provides unique obstacles due to the complexity of its morphology and syntax, and previously available pre-trained models sometimes needed to be revised to deal with these complexities. AraBERT was developed to address these insufficiencies in the existing models. AraBERT has achieved state-of-the-art results on several Arabic language benchmarks and has gained widespread adoption in the Arabic natural language processing (NLP) field for various applications. AraBERT's capacity to pick up on the subtleties of the Arabic language is one of its essential qualities. As a result, researchers and practitioners who engage with Arabic text data will find it a beneficial tool. The creation of AraBERT is a big step

forward in the processing of the Arabic language, and it has made new avenues available for the comprehension and use of the Arabic language in communication.

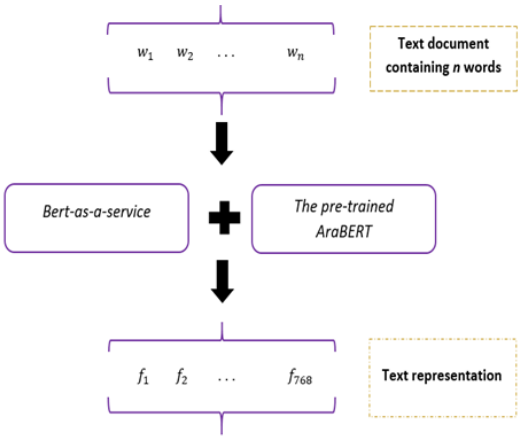

Figure 2. AraBERT feature extraction

mBERT

mBERT [19], "multilingual BERT," is a pre-trained deep learning model built by Google. It can comprehend and analyze text in a variety of languages. It is educated using a sizable corpus of text data derived from over one hundred different languages, including low-resource languages. mBERT has already been pre-trained on a language modeling task, and it can be fine-tuned on various NLP tasks, including sentiment analysis, named entity identification, and machine translation in several different languages. A multilingual model such as mBERT, which can be especially useful for low-resource languages requiring more training data or models already pre-trained, can be very effective. The NLP community has widely accepted it for cross-lingual transfer learning, and mBERT has obtained state-of-the-art performance on various cross-lingual benchmarks. Because of the development of mBERT, new opportunities have arisen for language comprehension and communication across various languages. This has enormous repercussions for cross-cultural communication and globalization.

B. DEEP LEARNING MODELS

Deep learning models are artificial neural network models that use numerous layers of nonlinear transformations to learn from big datasets. These models are utilized in various applications, such as computer vision, natural language processing, speech recognition, and recommendation systems. Our study used RNN, LSTM, BiLSTM, and TreeLSTM.

RNN

RNNs are deep learning models that analyze sequential data like text or audio. RNNs, unlike standard feedforward neural networks, can incorporate the context of past inputs when processing new inputs. As a result, they are precious for tasks like language modeling, speech recognition, and machine translation. RNNs function by transferring the output of a one-time step as input to the next time step, resulting in a loop that allows the network to keep an internal memory of past inputs. This enables the network to learn long-term dependencies between input data, which is difficult for other forms of neural networks to achieve [20].

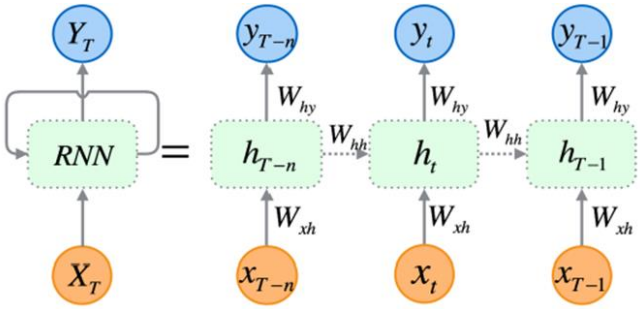

Figure 3. RNN structure

LSTM

Long Short-Term Memory (LSTM) networks are a Recurrent Neural Network (RNN) type that was made to deal with vanishing gradients in standard RNNs. They are especially good at handling sequential data and have been used extensively for jobs like speech recognition, machine translation, and image captioning. LSTMs have a unique design with a cell state that can be changed by gates controlling the information flow. These gates let the network choose which information from earlier inputs to remember or forget. This lets it deal with long-term dependencies in sequential data. LSTMs have succeeded in many areas, such as speech recognition, machine translation, and handwriting recognition. They are now seen as one of the most important deep learning architectures and have made a big difference in natural language processing [20].

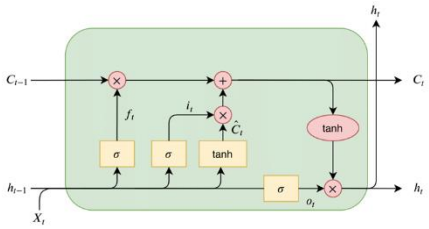

Figure 4. LSTM structure

BILSTM

Bidirectional Long Short-Term Memory (BiLSTM) is a variant of the Long Short-Term Memory (LSTM) network that has gained popularity in natural language processing tasks, including machine translation, named entity recognition, and sentiment analysis.

BiLSTMs function by processing input sequences in both forward and reverse orientations, enabling the network to capture information from both past and future inputs. This enables the network to manage long-term dependencies better and capture context-based data from the entire sequence. Like LSTMs, BiLSTMs use gates to remember or forget information from prior inputs selectively. However, they have two sets of gates, one for forward movement and one for reverse movement. This makes them particularly effective at tasks requiring understanding sentence context and word relationships. Overall, BiLSTMs have considerably enhanced the performance of natural language processing models, allowing for the creation of more precise and sophisticated applications [22].

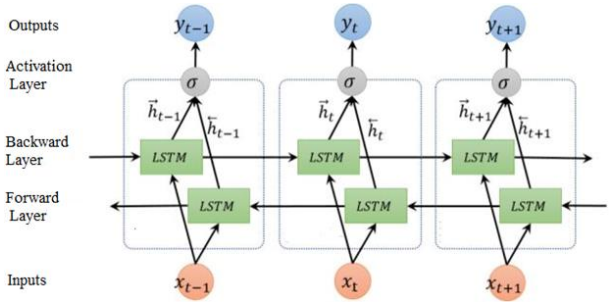

Figure 5. Bi-LSTM structure

TreeLSTM

Tree-structured Long Short-Term Memory (TreeLSTM) is a version of the Long Short-Term Memory (LSTM) network used to handle tree-structured data, like parse trees, in natural language processing. TreeLSTMs add to the basic LSTM design by giving each node in the tree structure a cell state and a way to let information in or out. This lets the network choose which information to keep or throw away at each node based on where it is in the tree. TreeLSTMs are very good at jobs like figuring out how someone feels about something by looking at the structure of the parse tree. They have also been used for adding captions to images and making code. Overall, TreeLSTMs have made it much easier for deep learning models to deal with ordered data like trees, graphs, and other hierarchical structures. This has given AI researchers new ways to work with natural language processing and other areas of AI [23].

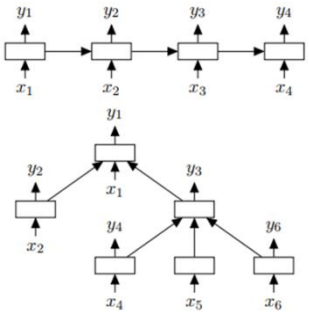

Figure 6. A tree-structured LSTM network with arbitrary branching factor

IV. EXPERIMENTS AND RESULTS

A. DATASET DESCRIPTION

Gathering textual data is the very first step in doing sentiment analysis. In order to evaluate how well our model performs in the context of this study, we looked at the following datasets:

Table 1. Experimental datasets

Name	Negative	Positive	Total
The Books Reviews in Arabic Dataset (BRAD) [24]	255,300	255,300	510,600
The Arabic Reviews dataset (ARD) [25]	50,000	50,000	100,000
Merged datasets (MD)	305,300	305,300	610,600

In order to prepare the data for the duties at hand, several cleansing procedures were performed. These comprised:

- The deletion of URLs and HTML elements.
- Elimination of all non-Arabic characters.
- Elimination of emoticons.
- The removal of repeated characters.
- Tokenization divides the text into smaller entities, such as words or sub words.
- Normalization and stemming involved converting words to their base form and reducing them to their root form.

These measures were taken to ensure the data for the tasks were accurate and properly formatted, making them simpler to analyze and process with machine learning algorithms. By removing extraneous information, such as URLs and non-Arabic characters, and reducing words to their basic form, the resulting data were more standardized and consistent, enhancing the accuracy and efficiency of subsequent analysis.

B. PERFORMANCE MEASURES

Several distinct metrics were used to evaluate the efficacy of the proposed model performance enhancement. The choice of metric can significantly impact the ability to monitor and compare the efficiency and effectiveness of various models. The evaluation of the quality of our research was based on five criteria, which are briefly summarized in Table 2. By utilizing multiple these metrics, we can better understand our model's performance and identify potential improvement areas. Accuracy, precision, recall, F1 score, and sensitivity may be among these criteria. By selecting and analyzing these metrics with care, we can more precisely evaluate the performance of our model and make informed decisions regarding future enhancements [26, 27].

Table 2. Description of evaluation metrics

Performance Evaluation	Summary	Equation
Accuracy	Accuracy measures how effectively a model predicts a task's outcome. It is calculated by dividing correct predictions by total predictions.	*(Tp+Tn)/(Tp+Tn+Fp+Fn)*
Precision	Precision is the percentage of correct positive forecasts. Dividing the number of true positives by the sum of true and false positives yields it.	*Tp/(Tp+Fp)*
F1-score	The F1 score combines precision and recall to evaluate a model. The harmonic mean of precision and recall.	*(2*(Precision.Recall))/((Precision+Recall))*
Recall	Recall is the percentage of true positive predictions among all positive samples. Divide the number of true positives by the sum of true positives and false negatives.	*Tp/(Tp+Fn)*
Specificity	Specificity measures how successfully a model detects negative samples. Divide the number of true negatives by the sum of true negatives and false positives.	*Tn/(Tn+Fp)*

C. EXPERIMENTAL RESULTS

In this section, we report the findings of our research analyzing the impact of various word embeddings on Arabic sentiment analysis using a variety of models, such as RNN, LSTM, BiLSTM, TreeLSTM, RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM. Specifically, we compared the outcomes of these analyses using RNN, LSTM, BiLSTM, and TreeLSTM. We used metrics such as accuracy, specificity, F1-score, precision, and recall for performance evaluation.

The findings from the several Arabic text datasets utilized are presented in Tables 3, 4, and 5. According to the findings, the suggested model, which uses a hybrid deep learning strategy, performs better than the base classifiers. It achieved the greatest accuracy rates on the BRAD, ARD, and MD datasets, with 92.03%, 92.58%, and 92.27%, respectively. This research shows that the proposed model is superior to the base classifiers. The suggested model displayed a considerable boost in accuracy for the MD dataset, with a 13.4% improvement compared to RNN, LSTM, BiLSTM, and TreeLSTM.

Table 3. Evaluation of the performance of implemented models in the BRAD dataset

BRAD dataset						
Deep learning techniques	***Word Embeddings***	***Precision***	***Accuracy***	***Recall***	***F1-measure***	***Specificity***
RNN	***BERT***	0,7221	0,7234	0,5478	0,7018	0,8157
	mBERT	0,6737	0,7821	0,5234	0,7145	0,8325
	AraBERT	0,7111	0,8155	0,6571	0,7412	0,8045
LSTM	***BERT***	0,6245	0,7258	0,4502	0,7042	0,8174
	AraBERT	0,6761	0,7854	0,5258	0,7169	0,8349
	mBERT	0,7135	0,8179	0,6595	0,7436	0,8064
BiLSTM	***BERT***	0,6563	0,7351	0,4479	0,6823	0,8374
	mBERT	0,7363	0,8443	0,5165	0,7357	0,7809
	AraBERT	0,7579	0,8626	0,6035	0,7682	0,8191
TreeLSTM	***BERT***	0,6971	0,7931	0,6543	0,7838	0,8154
	mBERT	0,7582	0,8106	0,7103	0,6793	0,8325
	AraBERT	0,8111	0,8893	0,6134	0,6153	0,8046
RNN-LSTM	***BERT***	0,6579	0,7482	0,6026	0,8126	0,8654
	mBERT	0,7925	0,8872	0,5474	0,7411	0,7511
	AraBERT	0,8233	0,9057	0,6426	0,8653	0,8041
RNN-BiLSTM	***BERT***	0,8149	0,7234	0,4478	0,7018	0,7651
	mBERT	0,6737	0,7821	0,5234	0,6145	0,8197
	AraBERT	0,7111	0,8155	0,6571	0,7412	0,7651
RNN-TreeLSTM	***BERT***	0,7079	0,8107	0,7251	0,7423	0,8051
	mBERT	0,7662	0,8621	0,5826	0,7455	0,6326
	AraBERT	**0,84516**	**0,9134**	**0,7355**	**0,8912**	**0,8843**

Table 4. Evaluation of the performance of implemented models in the BRAD dataset

ARD dataset						
Deep learning techniques	***Word Embeddings***	***Precision***	***Accuracy***	***Recall***	***F1-measure***	***Specificity***
RNN	***BERT***	0,7314	0,7328	0,5549	0,7109	0,8263
	mBERT	0,6824	0,7922	0,5302	0,7237	0,8433
	AraBERT	0,7203	0,8261	0,6656	0,7508	0,8149
LSTM	***BERT***	0,6326	0,7352	0,4560	0,7133	0,8280
	AraBERT	0,6848	0,7956	0,5326	0,7262	0,8457
	mBERT	0,7227	0,8285	0,6680	0,7532	0,8168
BiLSTM	***BERT***	0,6648	0,7446	0,4537	0,6911	0,8482
	mBERT	0,7458	0,8552	0,5232	0,7452	0,7910
	AraBERT	0,7677	0,8738	0,6113	0,7781	0,8297
TreeLSTM	***BERT***	0,7061	0,8034	0,6628	0,7939	0,8260
	mBERT	0,7680	0,8211	0,7195	0,6881	0,8433
	AraBERT	0,8216	0,9008	0,6213	0,6232	0,8150
RNN-LSTM	***BERT***	0,6664	0,7579	0,6104	0,8231	0,8766
	mBERT	0,8028	0,8987	0,5545	0,7507	0,7608
	AraBERT	0,8340	0,9174	0,6509	0,8765	0,8145
RNN-BiLSTM	***BERT***	0,8254	0,7328	0,4536	0,7109	0,7750
	mBERT	0,6824	0,7922	0,5302	0,6224	0,8303
	AraBERT	0,7203	0,8261	0,6656	0,7508	0,7750
RNN-TreeLSTM	***BERT***	0,7171	0,8212	0,7345	0,7519	0,8155
	mBERT	0,7761	0,8733	0,5901	0,7551	0,6408
	AraBERT	**0,85614**	**0,9252**	**0,7450**	**0,9027**	**0,8957**

Table 5. Evaluation of the performance of implemented models in the BRAD dataset

MD dataset						
Deep learning techniques	***Word Embeddings***	***Precision***	***Accuracy***	***Recall***	***F1-measure***	***Specificity***
RNN	***BERT***	0,7438	0,7452	0,5643	0,7229	0,8403
	mBERT	0,6940	0,8056	0,5392	0,7360	0,8576
	AraBERT	0,7325	0,8401	0,6769	0,7635	0,8287
LSTM	***BERT***	0,6433	0,7476	0,4637	0,7254	0,8420
	AraBERT	0,6964	0,8091	0,5416	0,7385	0,8600
	mBERT	0,7349	0,8425	0,6793	0,7660	0,8306
BiLSTM	***BERT***	0,6761	0,7572	0,4614	0,7028	0,8626
	mBERT	0,7584	0,8697	0,5320	0,7578	0,8044
	AraBERT	0,7807	0,8886	0,6216	0,7913	0,8438
TreeLSTM	***BERT***	0,7181	0,8170	0,6740	0,8073	0,8400
	mBERT	0,7810	0,8350	0,7317	0,6997	0,8576
	AraBERT	0,8355	0,9161	0,6318	0,6337	0,8288
RNN-LSTM	***BERT***	0,6777	0,7707	0,6207	0,8370	0,8915
	mBERT	0,8164	0,9139	0,5639	0,7634	0,7737
	AraBERT	0,8481	0,9329	0,6619	0,8914	0,8283
RNN-BiLSTM	***BERT***	0,8394	0,7452	0,4613	0,7229	0,7881
	mBERT	0,6940	0,8056	0,5392	0,6329	0,8444
	AraBERT	0,7325	0,8401	0,6769	0,7635	0,7881
RNN-TreeLSTM	***BERT***	0,7292	0,8351	0,7469	0,7646	0,8293
	mBERT	0,7892	0,8881	0,6001	0,7679	0,6516
	AraBERT	**0,8706**	**0,9409**	**0,7576**	**0,9180**	**0,9109**

When we examined the performance characteristics of the trained models using the BRAD dataset like shown in Table 3, we discovered that our proposed model, RNN-TreeLSTM, incorporating AraBERT word embedding, achieved the greatest accuracy with 91.34%. This was the case when we compared the models to one another. The specific findings are in Table 3, which can be seen here. In addition, the hybrid model RNN-LSTM with mBERT word embedding achieved the second-highest accuracy with a score of 90.57%.

These findings further prove that our suggested model, RNN-TreeLSTM, is effective in accurately classifying sentiment analysis for the BRAD dataset. The implementation of AraBERT word embedding significantly contributed to this endeavor's success, which resulted in an exceptionally high degree of precision. Using recursive neural networks (RNN) in conjunction with the TreeLSTM architecture successfully captured the sequential and hierarchical pattern of the text, which allowed for robust sentiment analysis.

In Table 4, various deep learning techniques, such as RNN, LSTM, BiLSTM, TreeLSTM, RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM, were evaluated in terms of their performance on sentiment analysis. In addition, various word embeddings, such as BERT, mBERT, and AraBERT, were used to evaluate the impact of these word embeddings on the accuracy of the models. The RNN-TreeLSTM model that uses AraBERT word embedding was able to attain the maximum level of accuracy, which was 92.52%.

Based on the findings, word embedding substantially impacts the models' performance; nonetheless, AraBERT performs consistently well across a variety of deep-learning techniques. These findings offer academics and practitioners interest in improving the accuracy of sentiment analysis performed on Arabic text by utilizing deep learning algorithms and word embeddings with significant insights.

Based on the data presented in Table 5, it is evident that our hybrid deep learning model, which incorporates AraBERT word embedding, attained an impressive 94.09 percent accuracy. Closely trailing, the hybrid model RNN-LSTM with AraBERT embedding demonstrated an impressive accuracy of 91.61 %.

These results demonstrate the efficacy of our hybrid deep learning method for accurately classifying sentiment analysis. Using the robust capabilities of AraBERT word embedding, our model could capture and comprehend the nuances of the text data, resulting in superior accuracy.

Our hybrid deep learning model's impressive accuracy highlights its potential for accurate sentiment analysis. By outperforming other models, it demonstrated the significance of using state-of-the-art word embeddings with advanced neural network architectures.

These findings provide valuable insight into the efficacy of various sentiment analysis models and word embeddings. They demonstrate the potential of our hybrid deep learning model with AraBERT word embedding as a potent instrument for accurately categorizing sentiment across various applications.Researchers and practitioners can use these findings to improve their sentiment analysis methodologies and obtain more precise results, especially when working with Arabic text data.

D. DISCUSSION AND CONTRIBUTION

The outcomes derived from the suggested methodology for sentiment analysis in Arabic are significant, as they have potential ramifications for several sectors and fields, particularly those focused on understanding public opinion and evaluating customer satisfaction. The approach's remarkable precision and efficiency make it a valuable tool for corporations, governments, and organizations that aim to extract significant information from user-generated content across various languages and themes.

This research presents an innovative approach to conducting sentiment analysis in online social media conversations by combining deep learning methodologies. The experimental findings demonstrate significant improvements in accuracy, precision, F1-measure, specificity, and recall compared to previous techniques. The methods employed in

this study show robust performance when applied to Arabic social media data.

Our research has substantially advanced the disciplines of natural language processing and information management in general. In this study, we propose a novel strategy for improving the robustness and efficacy of sentiment analysis in Arabic social media.

The article introduces a proposed approach that employs hybrid deep learning techniques for opinion mining. This approach offers significant practical implications for businesses and organizations relying extensively on social media analysis to improve their performance and enhance consumer satisfaction. The methodology employed in our study facilitates the ability of enterprises to make informed decisions based on data analysis. It enables examining patterns and trends in public sentiment across multiple languages and domains to enhance their products and services. The proposed methodology incorporates various deep learning techniques, such as RNN, LSTM, BiLSTM, TreeLSTM, RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM,

Furthermore, the integration of AraBERT architecture is implemented in the word embedding. This methodology can be easily implemented for polarity classification in various social media platforms and languages. Our process provides advantages for companies functioning within Arabic surroundings or serving a global clientele. This study provides valuable insights into the perspectives and attitudes of customers in various regions.

Additionally, our methodology can aid firms in surveilling and controlling their brand reputation on social media platforms by analyzing client attitudes and emotions. This approach can help firms identify possible concerns and quickly address client feedback, improving customer satisfaction and cultivating loyalty. In brief, our proposed methodology has substantial practical implications for enterprises and organizations aiming to leverage social media analysis to enhance consumer satisfaction and promote growth.

The hybrid deep learning methodology for Arabic opinion mining has shown promising results. However, it is essential to recognize that the performance of this method may be affected by the language and domain features of the social media data being analyzed. Despite the robust performance exhibited by our practices in terms of accuracy, precision, F1-measure, specificity, and recall, our research primarily focuses on analyzing social media discourses about diverse topics. The effectiveness of the suggested methodology may demonstrate variation when applied in different fields, such as politics, healthcare, or sports, because of potential differences in language usage and conversation patterns.

V. SUMMARY AND FUTURE DIRECTIONS

This study explores the difficulty of analyzing emotions expressed in Arabic text. We assessed the efficacy of an Arabic sentiment analysis system using hybrid deep learning models, such as RNN, LSTM, BiLSTM, TreeLSTM, RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM.

We conducted experiments on three comprehensive datasets, BRAD, ARD, and MD, to evaluate the efficacy of our proposed model. The results validated our model's applicability for analyzing sentiments in Arabic-language texts. Our methodology consisted of two principal stages. Initially, characteristics were extracted using the Arabert model. RNN, LSTM, BiLSTM, and TreeLSTM served as simple classifiers, while RNN-LSTM, RNN-BiLSTM, and RNN-TreeLSTM were hybrid deep learning models.

To validate our methodology, we used an original Arabic review dataset. Our recommended strategy outperformed the baseline models on the MD dataset when combined with Arabert, achieving an impressive accuracy of 0.9409. The study's findings cast light on the value of textual data for formulating strategies, enhancing competitiveness, and managing income for professionals.

Moving forward, we recognize the importance of accurately capturing the meaning of Arabic words, and we anticipate that addressing this limitation will result in additional performance enhancements. Therefore, we will attentively consider this aspect in our upcoming work.

References

[1] A. O. J. Ibitoye, and O. F.W. Onifade, "Utilizing RoBERTa Model for Churn Prediction through Clustered Contextual Conversation Opinion Mining," *Int. J. Intell. Syst. Appl*., vol. 15, no. 6, pp. 1–8, 2023, https://doi.org/10.5815/ijisa.2023.06.01.

[2] S. Sultana, S. Rahman Eva, N. Hasan Moon, A. Islam Jony, and D. Nandi, "A Comparison of Opinion Mining Algorithms by Using Product Review Data," *Int. J. Inf. Eng. Electron. Bus*., vol. 14, no. 4, pp. 28–38, 2022, https://doi.org/10.5815/ijieeb.2022.04.04.

[3] H. Elzayady, K. M. Badran, and G. I. Salama, "Arabic Opinion Mining Using Combined CNN – LSTM Models," *Int. J. Intell. Syst. Appl*., vol. 12, no. 4, pp. 25–36, 2020, https://doi.org/10.5815/ijisa.2020.04.03.

[4] N. Hicham, S. Karim, N. Habbat, "Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach," *IJECE*, vol. 13, no 4, pp. 4504, 2023, https://doi.org/10.11591/ijece.v13i4.pp4504-4515.

[5] N. Hicham, S. Karim, N. Habbat, "Enhancing Arabic sentiment analysis in e-commerce reviews on social media through a stacked ensemble deep learning approach," *MMEP*, vol. 10, no 3, pp. 790-798, 2023, https://doi.org/10.18280/mmep.100308.

[6] N. Hicham, S. Karim, and N. Habbat, "An efficient approach for improving customer sentiment analysis in the Arabic language using an Ensemble machine learning technique," *Proceedings of the 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet)*, 2022, pp. 1–6. https://doi.org/10.1109/CommNet56067.2022.9993924.

[7] A. Al-Hashedi et al., "Ensemble classifiers for Arabic sentiment analysis of social network (Twitter data) towards COVID-19-related conspiracy theories," *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1-10, 2022, https://doi.org/10.1155/2022/6614730.

[8] G. Alwakid, T. Osman, M. E. Haj, S. Alanazi, M. Humayun, N. U. Sama, "MULDASA: Multifactor lexical sentiment analysis of social-media content in nonstandard Arabic social media," *Applied Sciences*, vol. 12, no. 8, pp. 3806, 2022, https://doi.org/10.3390/app12083806.

[9] S. Albahli, "Twitter sentiment analysis: An Arabic text mining approach based on COVID-19," *Front. Public Health*, vol. 10, pp. 966779, 2022, https://doi.org/10.3389/fpubh.2022.966779.

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

[11] A. Alharbi, M. Kalkatawi, M. Taileb, "Arabic sentiment analysis using deep learning and ensemble methods," *Arab J Sci Eng*, vol. 46, no 9, pp. 8913-8923, 2021, https://doi.org/10.1007/s13369-021-05475-0.

[12] A. Oussous, A. A. Lahcen, S. Belfkih, "Impact of text pre-processing and ensemble learning on Arabic sentiment analysis," *Proceedings of the 2nd International Conference on Networking, Information Systems & Security, NISS19*, Rabat, Morocco: ACM Press, 2019, pp. 1-9. https://doi.org/10.1145/3320326.3320399.

[13] S. Al-Saqqa, N. Obeid, A. Awajan, "Sentiment analysis for Arabic text using ensemble learning," *Proceedings of the 2018 IEEE/ACS 15th IEEE International Conference on Computer Systems and Applications (AICCSA)*, Aqaba, October 2018, pp. 1-7. https://doi.org/10.1109/AICCSA.2018.8612804.

[14] S. Al-Azani E.-S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text," *Procedia Computer Science*, vol. 109, pp. 359-366, 2017, https://doi.org/10.1016/j.procs.2017.05.365.

[15] C. Sitaula, A. Basnet, A. Mainali, T. B. Shahi, "Deep learning-based methods for sentiment analysis on Nepali COVID-19-related tweets,"

Computational Intelligence and Neuroscience, vol. 2021, pp. 1-11, 2021, https://doi.org/10.1155/2021/2158184.
[16] C. Sitaula, T. B. Shahi, "Multi-channel CNN to classify Nepali Covid-19 related tweets using hybrid features," 2022, https://doi.org/10.1007/s12652-023-04692-9.
[17] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, P. Duan, "Word Embeddings and Convolutional Neural Network for Arabic Sentiment Classification », p. 11.
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, 2019. [Online]. Available at: http://arxiv.org/abs/1810.04805.
[19] M. Artetxe, H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond," 2018, https://doi.org/10.1162/tacl_a_00288.
[20] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," 2018, doi: 10.48550/ARXIV.1808.03314.
[21] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no 8, pp. 1735-1780, 1997, https://doi.org/10.1162/neco.1997.9.8.1735.
[22] Z. Huang, W. Xu, K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, doi: 10.48550/ARXIV.1508.01991.
[23] K. S. Tai, R. Socher, C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," 2015, https://doi.org/10.3115/v1/P15-1150.
[24] A. Elnagar, O. Einea, "BRAD 1.0: Book reviews in Arabic dataset," *Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, 2016, p. 1-8. https://doi.org/10.1109/AICCSA.2016.7945800.
[25] "Arabic 100k reviews," [Online]. Available at: https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews
[26] M. A. Muslim et al., "New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning," *Intelligent Systems with Applications*, vol. 18, article 200204, 2023. https://doi.org/10.1016/j.iswa.2023.200204.
[27] M. M. Abdelgwad, T. H. A. Soliman, A. I. Taloba, and M. F. Farghaly, "Arabic aspect based sentiment analysis using bidirectional GRU based models," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 9, pp. 6652–6662, 2022, https://doi.org/10.1016/j.jksuci.2021.08.030.

Nouri Hicham is a doctoral candidate at Hassan II University in Morocco. He has two master's degrees: one in spatial economy and territorial governance from the Faculty of Legal Economic and Social Sciences AIN SEBAA and another in data engineering from the Hassania School of Public Works. His primary areas of research interest are Artificial Intelligence and marketing.

Nassera HABBAT holds a Ph.D. from Hassan II University in Casablanca. 2015 she graduated from Caddi Ayyad University with a Master's in Information Systems Engineering. In addition, she obtained a Technology and Web Programming license from the same institution in 2013. Her research areas of interest include big data and machine learning.