A Thorough Analysis of E-commerce Customer Reviews in Arabic Language Using Deep Learning Techniques for Successful Marketing Decisions

Nouri Hicham, Habbat Nassera, and Sabri Karim

Abstract— **Recent years have seen the fast rise of technology related to the Internet. Online shopping has become an increasingly common option for people to purchase and fulfill their consumption needs. It can effectively boost consumer happiness by conducting sentiment analysis of many user evaluations on e-commerce platforms. This research proposes a new sentiment analysis model built on stacking hybrid deep learning models. In this research, we proposed a stacking technique that predicts accurately Arabic emotion by combining the prediction power of hybrid deep learning models such as RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiGRU-BiLSTM, and BiGRU-BiLSTM with MLP word embedding. The suggested model's effectiveness is evaluated using multiple open access datasets: AHS(sub), AHS(main), AR-twitter, LARGE, OCA, OCLAR, OD, and merged datasets with 104808 items. The experiments showed that the proposed model is appropriate for assessing the feelings in texts written in the Arabic language. The first stage of the proposed strategy consists of the feature extraction employing the Arabert algorithm. Next, We create and train five hybrid deep-learning models. The subsequent step is to concatenate the outputs of the fundamental classifiers with Multilayer Perceptron. Compared to conventional and hybrid deep learning approaches, our approach achieved an accuracy improvement of 0.9668.**

Index Terms— **Hybrid deep learning, Arabic language, sentiment analysis, classification, stacking**

I. Introduction

An ever-increasing number of customers choose to shop on different online shopping sites due to the rapid development and widespread adoption of technology that enables online commerce[1]. Customers may purchase online whenever and wherever they want, saving time and effort compared to traditional shopping. In addition, the products sold through e-commerce platforms come in a wide variety of styles, and customers may buy the things they want without leaving the comfort of their homes [1]. Because e-commerce platforms are virtual, various things could be improved with the things supplied[2]. These problems include descriptive material that does not match the actual goods and inadequate and poor-quality after-sales services [3]. Consequently, it is required to research on consumer opinions regarding the quality of commodities purchased through various e-commerce platforms.

Analyzing of the sentiment trend of consumer evaluations can serve as a resource for other consumers and help e-commerce platform firms improve the quality of their service and the level of pleasure their customers feel. A procedure that automatically examines the text of the subjective opinion with the customer's vibrant hue and extracts the customer's emotional disposition is called sentiment analysis for customer reviews. This process is called text-oriented analysis or opinion mining [4].

Throughout the previous decade, substantial advancements have been made in ASA. Constructing sentiment analysis systems for various domains has used multiple algorithms and machine learning approaches [5], The amount of research published in ASA has dramatically increased over the last few years. Naive Bayes and SVM algorithms account for more than 70% of all published efforts; nonetheless, shallow machine learning remains the most popular method in the articles analyzed by [4] [5]. Based on the same study, only 2% of ASA publications utilized deep learning (DL). This suggests that Arabic sentiment analysis is less advanced than analysis performed in other languages, where DL is quickly substitutes rudimentary machine learning.

In addition, most ASAs founded on deep learning trials utilized long short-term memory (LSTM) networks and convolutional neural networks (CNN) or a combination of the two. These sequential models need to be more adaptable to constructing random networks. In addition, the majority of the previous models relied on the word2vec embedding technique [7] or derivatives such as fastText [8]; these research techniques are not "completely contextualized," and they ignore the location and order of the words, which reduces the accuracy of the findings.

To find solutions to these problems, many researchers [9]–[11] have developed attention-based algorithms that

Manuscript received January 25, 2023; revised June 26, 2023.

Nouri Hicham is a PhD candidate of Research Laboratory on New Economy and Development (LARNED), Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University of Casablanca, Morocco. (Corresponding author e-mail: nhicham191@gmail.com).

Habbat Nassera is a PhD candidate of RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca, Morocco. (e-mail: nassera.habbat@gmail.com).

Sabri Karim is a professor of Research Laboratory on New Economy and Development (LARNED), Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University of Casablanca, Morocco. (e-mail: sabrikarimprof@gmail.com).

search for and magnify key context areas in input vectors. Most of these models are encoder-decoder models used for sequence-to-sequence data.

Consequently, we have developed an ensemble stacking model for ASA using Support Vector Machines (SVM) and Multilayer Perceptron (MLP) as the meta-learner. This model contains hybrid deep learning models composed of RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiLSTM-BiGRU, and BiGRU-BiLSTM, respectively. After completing the optimization technique, the created model's performance was finally the best compared to other hybrid models. This is a condensed version of the contributions that we have made, which can be stated as follows:

• For deep learning, we suggested utilizing hybrid architectures such as RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiLSTM-BiGRU, and BiGRU-BiLSTM.

• We present a model for ensemble stacking that incorporates the following hybrid DL models: RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiLSTM-BiGRU, and BiGRU-BiLSTM. The outputs of six hybrid deep learning bases have been combined using Multilayer Perceptron.

• The effectiveness of the stacking model was compared to the performance of numerous hybrid deep learning models using a variety of word embedding (AraVec, ELMO, AraBERT) models to determine whether the suggested ensemble model is superior in terms of recall, accuracy, f1-score, precision, and sensitivity. The suggested ensemble stacking approach considerably improved performance over previously developed deep learning models. The following outline constitutes our paper's structure: The second part of this chapter explains the accompanying sentiment analysis for Arabic models, which plays a crucial aspect in our work. The third section presents an outline of the suggested model for your consideration. The outcomes of the experiments are broken down and scrutinized in greater detail in Section 4. The final part of the article is the conclusion.

II. RELATED WORK AND LITERATURE REVIEW

Numerous attempts have been made to implement ML inside Arabic Sentiments Analysis using various techniques and models [8][9][11]. Traditional lexicon-based methods, shallowML, and DL models are some of the models used. Since the primary focus of this post is on DL and how it may be implemented to achieve cutting-edge outcomes in ASA, we will begin by defining deep learning. Many initiatives employing various conceptual frameworks and methodological techniques have been made to implement machine learning in ASA. The models utilized include time-honored lexicon techniques, shallow machine learning models, and, more recently, deep learning model applications. Since the primary focus of this post is on deep learning and how it might be applied to achieve state-of-the-art results in ASA, we will begin with some background information.

The researchers in [12] utilized character-level deep CNNs to analyze the sentiment of Arabic text. They assessed their model using a combined AS dataset of eight datasets containing more than 90,000 items. 94.33% accuracy was the greatest that could be accomplished. Their model had a 7% performance advantage over conventional machine learning classifiers.

The study's authors in [13] utilized a combined CNN-LSTM model for SA using the following datasets: Sub-AHS, Main-AHS, ASTD, and Ar-Twitter. They carried out a sentiment analysis on several levels, including the character level, the character N-gram level, and the word level. The research performed at the word level yielded positive results on some datasets.

In [14], the authors used a CNN with LSTM to assess sentiments in three publicly available datasets: ASTD, SemEval 2017, and ARSAS. To generate the necessary word embeddings, they used the word2vec model. Their investigations in categorization focused on the three different feelings that can be distinguished. Their model obtained results state-of-the-art results on the ASTD and SemEval 2017, providing an accuracy rate of 92% on the ARSAS dataset. The authors suggested utilizing a CNN-LSTM hybrid model with word2vec-derived features for the binary categorization of Arabic perspectives [8]; they used a variety of ASA datasets, such as Main-AHS, ASTD, and Ar-Twitter. The CNN-LSTM hybrid model achieved the highest level of accuracy, which was measured at 79.07%. Using CNN, the researchers [15] analyzed the binary sentiment of nine distinct datasets, including the ASTD and the LABR. Tweets and reviews comprise the two different categories that include the dataset. The Skip-Gram and the CBOW variants of word2vec were utilized in creating the word embedding matrix; They also tested CNN with datasets that differed significantly in terms of their level of balance.

The authors in [16] proposed a hybrid model that evaluates Arabic sentiment based on LSTM and RNN. This model is the LSTM-RNN hybrid model. They investigated the impact of utilizing Deep Learning with various pre-trained word embeddings on the results. They tested their model in a variety of environments with the AraSenTi-Tweet.

When using ensemble models, the power of inference can be improved using individual models. In addition, the efficiency of hybrid methods could be improved by implementing hybrid models as base classifiers within an ensemble [17]. Ensemble models have been utilized in various fields and have demonstrated superior performance to base models [18]. In the ASA field, an ensemble modeling technique has been applied. This is shown by the authors of [17], who developed an ensemble model that maximizes Arabic sentiment analysis via voting. Their methodology is called "voting to optimize." They used CNN-LSTM and an optimization strategy applied to the dataset of AS tweets (ASTD) to determine which LSTM and CNN were the most effective. The chosen models had the most significant possible f1 score compared to the other models. The authors of [17] presented an ensemble methodology for predicting the sentiment of tweets written in Arabic. This model, constructed using CNN and LSTM, was marketed as having the ability to make accurate

predictions. To gather the outcomes of the investigations, they utilized the dataset provided by ASTD. According to the findings, the ensemble model possesses the highest level of accuracy as well as the highest f1-score.

III. METHODOLOGY

We suggested a novel ASA method based on this investigation's stacked ensemble learning concept. The presented technique uses hybrid RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiLSTM-BiGRU, and BiGRU-BiLSTM classifiers. The proposed strategy combines the results of these hybrid models with those of meta-classifiers. By utilizing this strategy of stacked ensemble learning, we may simultaneously improve overall performance and capitalize on each model's structural and functional advantages.

A. Word Embedding(WE)

Word embedding is a group of machine learning approaches that aim to represent a text's words or sentences by vectors of real numbers represented in a Vector Space Model. These techniques are sometimes referred to together as word embedding. These novel representations of textual data have made it possible to increase the performance of artificial language processing approaches like Topic Modeling and Sentiment Analysis [19].

The linguistic theory, Distributional Semantics, developed by Zelling Harris, serves as the foundation for word embedding [20]. The surrounding circumstances in which a word is employed is one of the factors that are taken into consideration by this hypothesis. Consequently, the meanings of words are frequently interchangeable depending on the environment in which they are employed. Word embedding algorithms are used quite frequently to describe individual words through numerical vectors. However, these algorithms can also build vector representations of entire sentences, biological data such as DNA sequences, or networks represented as graphs [21]. Therefore, this section describes the various WE employed in this work.

AraVec

AraVec is an open-source program developed by [22] and offers robust word embedding models for Arabic natural language processing applications. AraVec provides six-word embedding models that have already been pretrained, and these models use data from Twitter, Wikipedia, and Common Crawl. The overall quantity of tokens that were utilized in the creation of the models is more remarkable than 3300000000. They have offered two models for every resource, one based on the CBOW model and the other on the SG model. These models were put through their paces on many tasks, including identifying word similarities and using qualitative and quantitative metrics to evaluate their performance. The findings that the suggested methodology produces are significant. An approach is a valuable tool for determining the degree of similarity between two words, and it also has the potential to enhance the effectiveness of other NLP tasks.

AraBERT

The Arabic pre-training BERT transformer model (AraBERT) expresses semantics in context through word embedding. It was trained on datasets taken from Arabic news websites. These datasets included 1 billion tokens spread across 3.5 million articles from the OSLAN Corpus and 1.5 billion words across 5 million articles from 10 primary news sources in 8 countries. The best-as-a-service technology is illustrated in Figure 1. This technique activates layers without requiring the settings of AraBERT to be fine-tuned [8]. It estimates the average value of the previous-to-last concealed token pool. The output representation is incorporated into the classification models that we shall talk about in the following sentence.

ELMO

ELMO is a deep contextualized word representation developed by [23]. It replicates the following:

- The difficulty of language usage, including syntax and semantics.

-In what ways do these usages differ depending on the context of the language used?

These word vectors can represent the deep bi-directional language model's (biLM) underlying state learning functions pre-trained on a considerable text corpus. They can be incorporated into the pre-existing models, which significantly advances the current state of the art across a wide range of challenging NLP applications, such as sentiment analysis, question answering, and textual entailment. All ELMO models were trained on the data from the 1 Billion Word Benchmark [24].

B. Deep learning Techniques

Recurrent neural network (RNN)

A recurrent neural network, often known as an RNN, is a form of architecture used in deep learning that processes sequential data. The connections between the neurons in an RNN form a directed graph. Within the context of this design, the internal state has been used to process the input sequence. Therefore, the plan is suitable for use in sequential processes, such as speech recognition, where it performs well. In RNN, each output was determined by repeatedly processing the same task over the sequence instances. This was done to find the outcomes. The result was obtained based on all of the computations that were done earlier [25]. As depicted in Figure 1, this structure gives context.

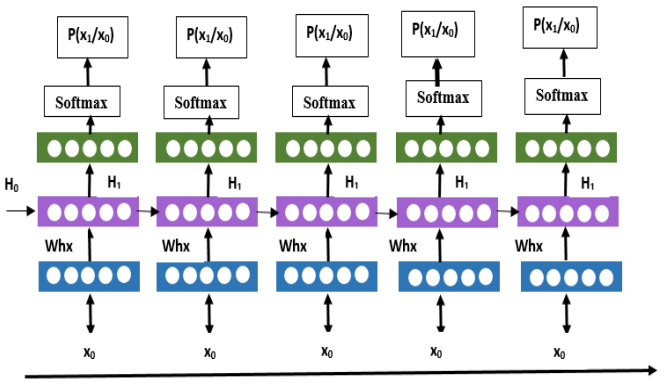

Fig. 1. RNN with Natural Language Mode

Bidirectional GRU (Bi-GRU)

The loss of information directly results from GRU networks' inability to exploit the present or future context. Data can be processed in both directions with the assistance of bidirectional GRU, also known as bi-GRU, which many researchers have utilized. All information from the hidden levels is brought up to the output layer. A GRU network that can communicate in both directions is created when two GRUs are connected. While the input sequence of one network is presented in the conventional chronological order, the series of another network is shown in the opposite direction. At each stage, the outputs of the two networks are brought together and combined [26]. The context can be seen in Figure 2.

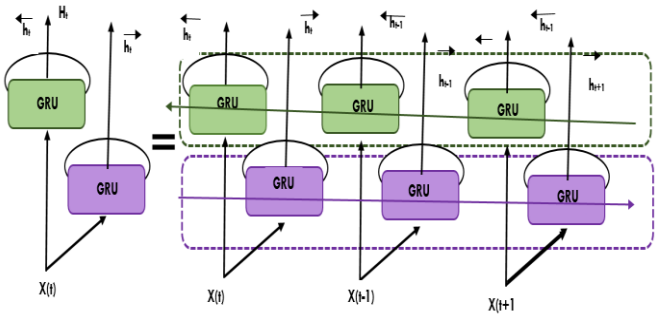

Fig. 2. Bi-GRU structure

Bidirectional LSTM (Bi-LSTM)

The LSTM network, which is more generic, has a subtype referred to as the BiLSTM network [27]. It utilizes a bidirectional network, which means that the inputs will cycle in two different directions, as illustrated in Figure 3, one from the future to the past and the other from the past to the end. This implies that the network can process information in both directions simultaneously. This allows the network to learn simultaneously in the forward and backward directions. As a consequence of this, it can maintain knowledge that is pertinent to both the past and the future at any one point in time by making use of the two hidden states simultaneously.

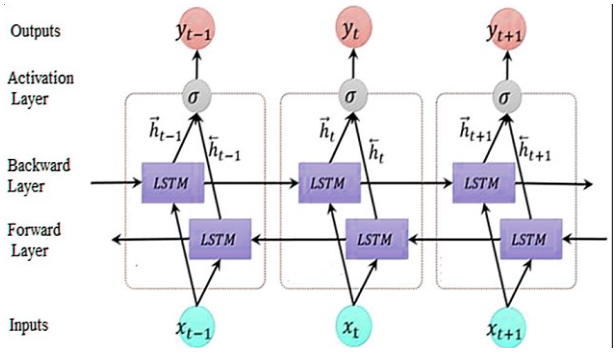

Fig. 3. Bi-LSTM structure

Hybrid Networks

The combination of RNN, Bi-GRU, and Bi-LSTM models brings together three distinct sorts of model types, each of which has its own set of advantages and unique architecture:

* Employed Bi-LSTM and Bi-GRU frameworks that would save data in the memory from the future while simultaneously utilizing both hidden states to protect data from the past as well as the future

* Utilized hidden states to save data from both the past and the future by utilizing the max-pooling layer; RNN is renowned for its capacity to extract the most significant possible number of features from the input data. This skill has helped RNN earn widespread recognition.

As shown in Figure 4, the input layer of both the Bi-LSTM and the Bi-GRU model commonly accepts vectors as input. These vectors come from different contextualized word embeddings. After that, the output of the Bi-LSTM and Bi-GRU models is used as the input for the RNN model. The purpose of combining these two architectural approaches is to produce a hybrid model that, in addition to the advantages provided by the Bi-LSTM and Bi-GRU models, can take advantage of the benefits offered by RNN. After each filter, a max pooling layer is applied to reduce the total amount of data while preserving the most current state of the system. The outputs of these final layers are concatenated and linked together before being integrated into a single two-dimensional softmax output. Ultimately, we classify datasets according to their polarity by applying the sigmoid activation function. This allows us to extract the anticipated outcome from the datasets. Using this function, binary labels can be applied to datasets.

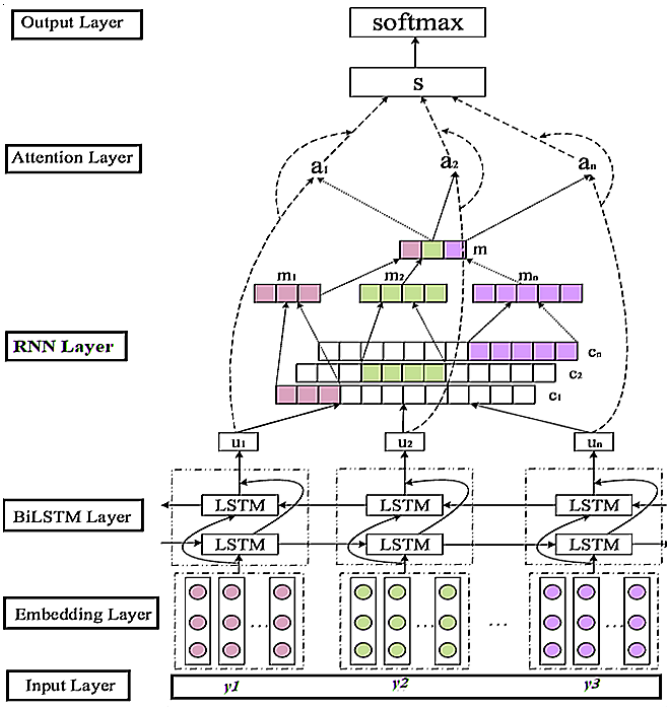

Fig. 4. BiLSTM-RNN architecture

IV. RESULTS AND EXPERIMENTS

In the next part, we will talk about the assessment metric we used to analyze the performance of our model, followed by our findings. After that, we shall share our concluding perspectives.

A. Dataset description

The datasets that were utilized in our study are outlined in Table 1. The table presents each set's primary characteristics, including the matching references, size,

category, number of positive items, and number of negative items.

The following preparatory procedures were carried out in order to cleanse the data and make them appropriate for the tasks in question:

- Deleting Removing URLs and HTML tags.
- Deleting non-Arabic characters.
- Deleting emojis.
- Letters that showed up more than twice were taken out.
- Tokenization.
- Normalization and stemming.

B. *Performance measures*

Several distinct indicators are utilized to assess how successfully the suggested enhancement to our model's performance has been implemented. The choice of metric can affect how the efficiency and effectiveness of different models are tracked and compared. Table 2 provides a concise description of the five criteria that served as the basis for evaluating the quality of our research[28]–[30].

TABLE II
A SUMMARY OF SENTIMENT ANALYSIS PERFORMANCE METRICS

Performance Evaluation	Summary	Equation
Accuracy(A)	Accuracy can be defined as the proportion of actual occurrences that correspond to correct predictions relative to the total number of actual occurrences .	$\frac{Tp+Tn}{Tp+Tn+Fp+Fn}$
Precision(P)	Precision can be defined as the proportion of samples that are correctly predicted in relation to the total number of samples.	$\frac{Tp}{Tp+Fp}$
F1-score(F)	The F1 score is calculated by finding the harmonic mean of the recall and precision scores. It is a number that combines accuracy and recall into a single unit.	$\frac{2*(\text{Precision}.\text{Recall})}{(\text{Precision}+\text{Reacall})}$
Specificity	Specificity is the ratio of the number of true negatives to the total number of negatives in the data.	$\frac{Tn}{Tn+Fp}$
Recall (R)	The recall is determined by taking the ratio of the total number of positive samples to the number of positive samples accurately categorized as positive.	$\frac{Tp}{Tp+Fn}$

C. *Experimental parameters*

The following were used as parameters in our experiments, as shown in Table 3:

D. *Experimental results*

In this part, we examine the results of our research that compared the effect of several different word embeddings on the classification of Arabic sentiment analysis carried out with RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiLSTM-BiGRU, BiGRU-LSTM, and our stacked model. We evaluated performance based on five metrics: accuracy, precision, the F1-score, recall, and sensitivity.

TABLE I
ARABIC SENTIMENT DATASETS

Name	Domain	Negative	Positive	Total
AHS(sub)[30]	Health	1231	502	1733
AHS(main) [30]	Health	1398	628	2026
Ar-twitter[31]	Tweets	1000	1000	2000
LARGE[32]	Multireviews	81	2073	2154
OCA[33]	Movie reviews	250	250	500
OCLAR[34]	Movies reviews	451	3465	3916
OD[14]	Mixed	21081	71398	92492
Merged datasets(MD)	Multidomain	25492	79316	104808

TABLE III
EXPERIMENTAL PARAMETERS

Configuring Selected Parameters	Score ranges	Best value
Batch size	80. 32. 64. 8.16	64
Epoch	30.15. 20	15
Optimizer	Adadelta, Adamax,Adagrad, Adadelta, Adam, Nadam	Adam
Dropout	0,6. 0,5. 0,4. 0,2	0,4
Activation function	softmax, softplus, relu, tanh, linear	softplus

In order to assess the efficacy of the stacked model, several experiments are carried out, each of which compares the proposed method to various other methodologies using a variety of open-source datasets. AHS(sub) [31], AHS(main) [31], AR-twitter[32], LARGE[33], OCA[34], OCLAR[35], OD[12], and merged datasets are all examples of the types of data sets utilized in this study. Table 1 provides specifics about the datasets combined by MD to create the stacked model used to evaluate our model's efficacy. Various datasets were included in the evaluation. The experiments utilize the configurations outlined in Table 3 of the experimental parameters.

Before moving on to the next step, we will apply our stacked model in isolation to every dataset in the binary categorization. This comparison aims to evaluate the accuracy of our results concerning the state-of-the-art accuracies obtained for each baseline in cases where those accuracies are available. Because all baseline models merely report the test accuracies for the datasets, comparisons with these modes are based on Our stacked model's test accuracies. Although most of these datasets are unbalanced, We determined the metrics for recall, precision, F1score, and accuracy to validate the dependability of the obtained results. Table 8 contains an inventory of the accuracies achieved by our stacked models when applied to the binary classification datasets. After this step, our stacked model is applied to the combined datasets, and the results for the combined dataset are presented in Tables 6 and 8.

As seen in Table 4, our stacked model, consisting of an MLP meta-learner and an AraBERT word embedding, achieved the highest accuracy, 94.78%, followed by the SVM meta-learner model, which achieved 94.61%. The

RNN-BiLSTM model that employed the AraVec word embedding achieved the lowest level of accuracy, reaching a maximum of 72.48 percent.

Concerning the other evaluation metrics, which are precision, F1-score, and recall, our model, which stacked hybrids of deep learning models, RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiGRU-BiLSTM, and BiGRU-BiLSTM with the word embedding MLP, presents good results in all of these metrics, which are 96.41%. 83,57%. 91,74% in the OD dataset, and 97,08%. 84,15%. 92.38 percent, respectively, in the AR-Twitter dataset, and for Merged datasets (MD) in Table 6, a performance that rises by 96.34% is presented. 85,24%. 93.58 percent in precision, 93.58 percent in recall, and 93.58 percent in F1-score, respectively.

In terms of sensitivity evaluation metric, as seen in Table 7, our stacked model, consisting of an MLP meta-learner and an AraBERT word embedding, achieved the highest sensitivity level with the value that rose 91.81%, 89.12% and 84.79% in the OD dataset, Ar-twitter dataset, and merged datasets used respectively.

TABLE IV
STACKED MODEL PERFORMANCE EVALUATION IN THE OD DATASET

OD Dataset[14]					
Deep learning techniques	***WE***	***A (%)***	***P (%)***	***R (%)***	***F (%)***
RNN-BiLSTM	***AraVec***	72,48	77,41	54,80	72,81
	ELMO	77,64	83,28	62,36	74,07
	AraBERT	81,38	86,62	75,73	76,74
BiLSTM-RNN	***AraVec***	75,90	78,57	54,81	70,85
	ELMO	83,90	89,50	61,67	76,19
	AraBERT	86,06	81,41	70,37	79,44
RNN-BiGRU	***AraVec***	79,98	84,38	75,45	81,07
	ELMO	86,09	80,89	81,05	70,55
	AraBERT	91,38	93,97	71,36	64,15
BiGRU-RNN	***AraVec***	76,06	79,89	70,28	83,88
	ELMO	89,52	93,79	64,76	76,73
	AraBERT	92,67	95,64	74,28	89,15
BiLSTM-BiGRU	***AraVec***	91,76	77,41	54,80	72,84
	ELMO	77,64	83,28	62,36	64,07
	AraBERT	81,38	86,62	75,73	76,74
BiGRU-LSTM	***AraVec***	84,17	85,09	74,84	79,19
	ELMO	83,75	82,39	71,71	78,54
	AraBERT	84,02	82,82	75,18	82,92
Our Stacked model with MLP	***AraVec***	81,06	86,14	82,53	76,85
	ELMO	86,89	91,28	68,28	77,17
	AraBERT	**94,78**	**96,41**	**83,57**	**91,74**
Our Stacked model with SVM	***AraVec***	80,89	85,97	82,36	76,68
	ELMO	86,72	91,11	68,11	77,89
	AraBERT	94,61	96,24	83,40	91,57

TABLE V
STACKED MODEL PERFORMANCE EVALUATION IN THE AR-TWITTER DATASET

Ar-twitter[31]					
Deep learning techniques	***WE***	***A (%)***	***P (%)***	***R (%)***	***F (%)***
RNN-BiLSTM	***AraVec***	72,98	77,95	55,18	73,30
	ELMO	78,18	83,86	62,79	74,58
	AraBERT	81,94	87,22	76,26	77,27
BiLSTM-RNN	***AraVec***	76,43	79,11	55,19	71,34
	ELMO	84,48	90,12	62,10	76,72
	AraBERT	86,66	81,48	70,86	79,99
RNN-BiGRU	***AraVec***	80,53	84,97	75,97	81,63
	ELMO	86,69	80,96	81,61	71,04
	AraBERT	92,01	94,62	71,85	64,59
BiGRU-RNN	***AraVec***	76,59	80,44	70,77	84,46
	ELMO	90,14	94,44	65,21	77,26
	AraBERT	93,31	96,30	74,79	89,77
BiLSTM-BiGRU	***AraVec***	92,40	77,95	55,18	73,34
	ELMO	78,18	83,86	62,79	64,51
	AraBERT	81,94	87,22	76,26	77,27
BiGRU-LSTM	***AraVec***	84,75	85,68	75,36	79,74
	ELMO	84,33	82,96	72,21	79,08
	AraBERT	84,60	83,39	75,70	83,50
Our Stacked model with MLP	***AraVec***	81,62	86,74	83,10	77,38
	ELMO	87,49	91,91	68,75	77,71
	AraBERT	**95,44**	**97,08**	**84,15**	**92,38**
Our Stacked model with SVM	***AraVec***	81,45	86,57	82,93	77,21
	ELMO	87,32	91,74	68,58	78,43
	AraBERT	95,27	96,91	83,97	92,21

TABLE VI
STACKED MODEL PERFORMANCE EVALUATION IN THE MERGED DATASETS (MD) DATASET

Merged datasets(MD					
Deep learning techniques	***WE***	***A (%)***	***P (%)***	***R (%)***	***F (%)***
RNN-BiLSTM	***AraVec***	73,92	78,96	55,89	74,25
	ELMO	79,19	84,95	63,60	75,54
	AraBERT	83,00	88,35	77,25	78,27
BiLSTM-RNN	***AraVec***	77,42	80,13	55,90	72,26
	ELMO	85,57	91,29	62,90	77,71
	AraBERT	87,78	82,53	71,78	81,02
RNN-BiGRU	***AraVec***	81,57	86,07	76,95	82,69
	ELMO	87,81	82,01	82,67	71,96
	AraBERT	93,20	95,85	72,78	65,42
BiGRU-RNN	***AraVec***	77,58	81,48	71,69	85,55
	ELMO	91,31	95,66	66,05	78,26
	AraBERT	94,52	97,55	75,76	90,93
BiLSTM-BiGRU	***AraVec***	93,60	78,96	55,89	74,29
	ELMO	79,19	84,95	63,60	65,34
	AraBERT	83,10	88,35	77,25	78,27
BiGRU-LSTM	***AraVec***	85,85	86,79	76,33	80,77
	ELMO	85,42	84,03	73,14	80,10
	AraBERT	85,69	84,47	76,68	84,58
Our Stacked model with MLP	***AraVec***	82,68	87,86	84,18	78,38
	ELMO	88,62	93,10	69,64	78,72
	AraBERT	**96,68**	**96,34**	**85,24**	**93,58**
Our Stacked model with SVM	***AraVec***	82,53	87,69	84,00	78,21
	ELMO	88,45	92,93	69,47	79,44
	AraBERT	96,57	95,16	85,06	93,40

TABLE VII
STACKED MODEL SPECIFICITY PERFORMANCE EVALUATION (%)

Deep learning techniques	*WE*	*OD Dataset*	*Ar-twitter*	*Merged datasets*
RNN-BiLSTM	***AraVec***	70,19	73,04	55,59
	ELMO	75,23	78,58	63,26
	AraBERT	78,82	81,72	76,84
BiLSTM-RNN	***AraVec***	73,52	74,12	55,61
	ELMO	81,26	84,44	62,56
	AraBERT	83,36	76,34	71,43
RNN-BiGRU	***AraVec***	77,46	79,62	76,54
	ELMO	83,39	75,86	82,23
	AraBERT	88,50	88,66	72,39
BiGRU-RNN	***AraVec***	73,67	75,37	71,31
	ELMO	86,71	88,49	65,75
	AraBERT	89,76	90,24	75,36
BiLSTM-BiGRU	***AraVec***	88,88	73,04	55,59
	ELMO	75,20	78,58	63,26
	AraBERT	78,91	81,72	76,84
BiGRU-LSTM	***AraVec***	81,52	80,28	75,92
	ELMO	81,12	77,73	72,75
	AraBERT	81,37	78,14	76,27
Our Stacked model with MLP	***AraVec***	78,51	81,27	83,73
	ELMO	84,15	86,12	69,27
	AraBERT	**91,81**	**89,12**	**84,79**
Our Stacked model with SVM	***AraVec***	78,37	81,11	83,55
	ELMO	83,99	85,96	69,10
	AraBERT	91,70	88,02	84,61

TABLE VIII
STACKED MODEL PERFORMANCE EVALUATION IN THE ARABIC SENTIMENT DATASETS

Name	*A (%)*	*P (%)*	*R (%)*	*F (%)*
AHS(sub)[30]	89,96	89,89	90,28	93,88
AHS(main) [30]	91,52	93,79	87,76	90,73
Ar-twitter[31]	95,44	97,08	84,15	92,38
LARGE[32]	92,6	95,64	89,28	91,95
OCA[33]	91,76	90,41	87,8	92,8
OCLAR[34]	92,64	93,28	89,36	94,07
OD[14]	94,78	96,41	83,57	91,74
Merged datasets(MD)	96,68	96,34	85,24	93,58

E. Use case

In this part, we will implement the proposed model (AraBERT + stacked hybrid deep learning model using MLP classifier) in a real dataset. This dataset comprises 93 147 Arabic reviews crawled from the Twitter page concerning Morocco Telecom between 07 September 2022 and 14 November 2022. Established in 1998 due to the demerger of the National Post and Telecommunications Office, it is the country's incumbent operator and, to this day, the most successful global telecommunications provider in Morocco. This collection concerned six cities in Morocco; Casablanca, Marrakech, Fez, Agadir, Tanger, and Oujda.

In order to compile this dataset, we pulled data from Twitter by utilizing the Twitter Application Programming Interface (API), which the Twitter Platform supplies. After registering for an account on https://apps.twitter.com, we were granted permission to access the database by utilizing four secret keys (consumer key, secret consumer key, access token, and secret access token) and gather tweets by utilizing the REST API.

We can obtain tweets written in Arabic about Morocco by filtering tweets according to place and language. We used the Tweepy library in Python to manage this data, and the Pymongo module in Python to store the data that collected in a MongoDB database. Figure 5 illustrates the sentiment analysis results in the Arabic language in these six different Moroccan cities.

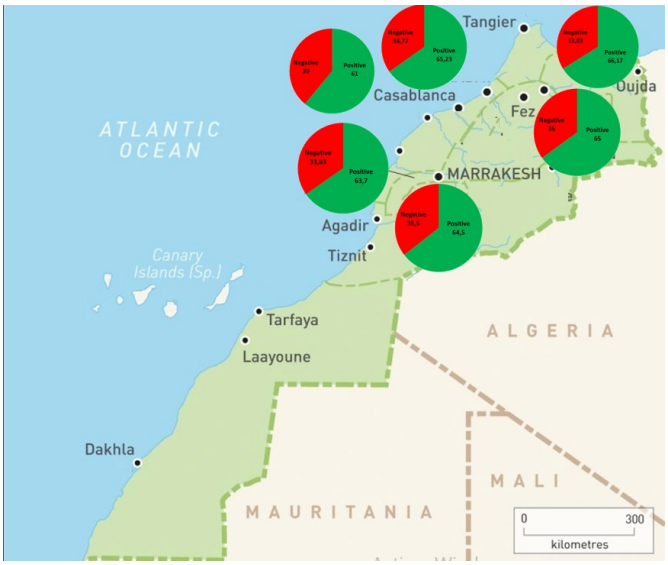

Fig.5. Arabic Sentiment analysis using the proposed model on real dataset by Moroccan city.

V. CONCLUSION AND FUTURE SCOPE

Novel deep learning based on stacked hybrids model; RNN-BiLSTM, BiLSTM-RNN, RNN-BiGRU, BiGRU-RNN, BiGRU-BiLSTM, and BiGRU-BiLSTM with MLP word embedding has been utilized to improve ASA using multiple open access public datasets: AHS(sub), AHS(main), AR-twitter, LARGE, OCA, OCLAR, OD, and merged datasets with 104808 items. The accuracies attained are 96,68%, 95,44%, and 94,78%, respectively, in the MD, Ar-twitter, and OD datasets. This model has a great deal of promise and potential. Our model beat all known attention models at the universal level, attaining top ranks on the leaderboards of all datasets utilized. A more significant amount of effort is required to build larger datasets for ASA and to further refine and optimize this model as a new benchmark strategy for multilingual and Arabic sentiment research.

REFERENCES

[1] R. Liang and J. Wang, "A Linguistic Intuitionistic Cloud Decision Support Model with Sentiment Analysis for Product Selection in E-commerce ", *Int. J. Fuzzy Syst*., vol. 21, no. 3, pp. 963-977, 2019.

[2] J.-R. Chang, L.-S. Chen, and L.-W. Lin, " A Novel Cluster based Over-sampling Approach for Classifying Imbalanced Sentiment Data ", *IAENG International Journal of Computer Science*, vol. 48, no. 4,pp. 1118-1128, 2021.

[3] P. Ji, H.-Y. Zhang, and J.-Q. Wang, " A Fuzzy Decision Support Model With Sentiment Analysis for Items Comparison in e-Commerce: The Case Study of http://PConline.com ", *IEEE Trans. Syst. Man Cybern, Syst*., vol. 49, no. 10, pp. 1993-2004, 2019.

[4] D. Zeng, Y. Dai, F. Li, J. Wang, and A. K. Sangaiah, "Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism ", *IFS*, vol. 36, no. 5, pp. 3971-3980, 2019.

[5] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, " A comprehensive survey of arabic sentiment analysis ", *Information Processing & Management*, vol. 56, no. 2, pp. 320-342, 2019.

[6] M. Wankhade, A. C. S. Rao, and C. Kulkarni, " A survey on sentiment analysis methods, applications, and challenges ", *Artif Intell Rev*, vol. 55, no. 7, pp. 5731-5780, 2022.

[7] N. Habbat, H. Anoun, and L. Hassouni, " Sentiment Analysis and Topic Modeling on Arabic Twitter Data during Covid-19 Pandemic ", *IJIAS*, vol. 2, no. 1, pp. 60-67, 2022.

[8] N. Habbat, H. Anoun, and L. Hassouni, " A Novel Hybrid Network for Arabic Sentiment Analysis using fine-tuned AraBERT model ",*International Journal on Electrical Engineering and Informatics*, vol. 13, no. 4, pp. 801-812, 2022.

[9] N. Hicham and S. Karim, " 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*, vol. 13, no. 10, pp. 122-130, 2022.

[10] 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 ", in 2022 *5th International Conference on Advanced Communication Technologies and Networking (CommNet),* 2022, pp. 1-6.

[11] I. Suliani, H. Asnal, L. Suryati, and R. Efendi, " Sentiment Analysis for Abolition of National Exams in Indonesia using Support Vector Machine ", *Engineering Letters*, vol. 30, no. 4,pp. 1342-1352, 2022.

[12] E. Omara, M. Mosa, and N. Ismail, " Deep Convolutional Network for Arabic Sentiment Analysis ", in 2018 *International Japan-Africa Conference on Electronics, Communications and Computations (JAC-ECC)*, Alexandria, Egypt: IEEE, 2018, pp. 155-159.

[13] A. M. Alayba, V. Palade, M. England, and R. Iqbal, " A Combined CNN and LSTM Model for Arabic Sentiment Analysis ", in 2018 *Cross Domain Conference for Machine Learning and Knowledge Extraction* , 2018, pp. 179-191.

[14] I. Abu Farha and W. Magdy, " Mazajak: An Online Arabic Sentiment Analyser ", in Proceedings of the *Fourth Arabic Natural Language Processing Workshop*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 192-198.

[15] N. Hicham, S. Karim, and 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.

[16] M. Heikal and M. Torki, " Sentiment Analysis of Arabic Tweets using Deep Learning ", in *2018 The 4th International Conference on Arabic Computational Linguistics (ACLing)*. 2018, pp. 114-12.

[17] H. Saleh, S. Mostafa, A. Alharbi, S. El-Sappagh, and T. Alkhalifah, " Heterogeneous Ensemble Deep Learning Model for Enhanced Arabic Sentiment Analysis ", *Sensors*, vol. 22, no. 10, pp. 3707-3735, 2022,

[18] M. Kang, J. Ahn, and K. Lee, " Opinion mining using ensemble text hidden Markov models for text classification ", *Expert Systems with Applications*, vol. 94,no. 15, pp. 218-227, 2018.

[19] R. Feng, C. Yang, and Y. Qu, " A Word Embedding Model for Analyzing Patterns and Their Distributional Semantics ", *Journal of Quantitative Linguistics*, vol. 29, no 1, pp. 80-105, 2022.

[20] G. Boleda, " Distributional Semantics and Linguistic Theory ", *Annu. Rev. Linguist*., vol. 6, no 1, pp. 213-234, 2020.

[21] E. Pavlick, " Semantic Structure in Deep Learning ", *Annual Review of Linguistics*, vol. 8, no 1, pp. 447-471, 2022.

[22] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, " AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP ", in *2017 International Conference on Arabic Computational Linguistics*, 2017, pp. 256-265.

[23] M. E. Peters and al. M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, " Deep contextualized word representations ", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.2018, pp. 2227–2237.

[24] C. Chelba and al., " One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling ", in *INTERSPEECH 2014* . 2014, pp. 2635-2639.

[25] L. Zhang, S. Wang, and B. Liu, " Deep learning for sentiment analysis: A survey ", *WIREs Data Mining Knowl Discov*, vol. 8, no. 4,pp,1253-1287, 2018.

[26] 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.

[27] M. Schuster and K. K. Paliwal, " Bidirectional recurrent neural networks ", *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[28] N. Hicham and S. Karim, " Machine Learning Applications for Consumer Behavior Prediction ", *Lecture Notes in Networks and Systems: Proceedings of Innovations in Smart Cities Applications 2022*, 2022. pp. 666-675.

[29] E. Bulbul, A. Cetin, and I. A. Dogru, " Human Activity Recognition Using Smartphones ", in 2018 2nd *International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara: IEEE, 2018, pp. 1-6.

[30] Cong-Cuong Le, P.W.C. Prasad, Abeer Alsadoon, L. Pham, and A. Elchouemi, " Text Classification: Naïve Bayes Classifier with Sentiment Lexicon " *IAENG International Journal of Computer Science*, vol. 46, no. 2, pp. 141-148, 2019.

[31] A. M. Alayba, V. Palade, M. England, and R. Iqbal, " Arabic language sentiment analysis on health services ", 2017 *1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*,2017, pp. 114-118.

[32] N. Hicham, S. Karim, and N. Habbat, " Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach ", *IJECE*, vol. 13, no. 4, pp. 4504-4515, 2023.

[33] H. ElSahar and S. R. El-Beltagy, " Building Large Arabic Multi-domain Resources for Sentiment Analysis ", in *Conference on Intelligent Text Processing and Computational Linguistics*, 2015, pp 23–34.

[34] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. U. López, et J. M. P. Ortega, " OCA: Opinion corpus for Arabic ", *Journal of the American Society for Information Science and Technology*, vol. 62,no. 10, pp. 2045-2054, 2011.

[35] M. A. Omari, M. Al-Hajj, N. E. Hammami, and A. Sabra, " Sentiment Classifier: Logistic Regression for Arabic Services' Reviews in Lebanon ", 2019 *International Conference on Computer and Information Sciences (ICCIS*),2019, pp. 1-5.

NOURI Hicham is a doctoral student at Hassan II University in Morocco. In 2021, he obtained from Faculty of Legal Economic and Social Sciences AIN SEBAA a master's degree in spatial economy and territorial governance, and another master's degree in data engineering from Hassania School of Public Works.

My research areas are Artificial Intelligence and marketing. He can be contacted at email: nourihicham@ieee.org/nhicham191@gmail.com.

Memberships:

- IAENG membership
- IEEE membership
- Institute for engineering research and publication membership