International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 4, August 2023, pp. 4504~4515
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i4.pp4504-4515

Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach

Nouri Hicham[1], Sabri Karim[1], Nassera Habbat[2]
[1]Research Laboratory on New Economy and Development, Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University of Casablanca, Casablanca, Morocco
[2]RITM Laboratory, CED ENSEM Ecole Superieure de Technologie, Hassan II University, Casablanca, Morocco

Article Info

Article history:

Received Oct 21, 2022
Revised Dec 19, 2022
Accepted Jan 2, 2023

Keywords:

Arabic language
Ensemble classifier
Machine learning
Novel approach
Sentiment analysis

ABSTRACT

Arabic's complex morphology, orthography, and dialects make sentiment analysis difficult. This activity makes it harder to extract text attributes from short conversations to evaluate tone. Analyzing and judging a person's emotional state is complex. Due to these issues, interpreting sentiments accurately and identifying polarity may take much work. Sentiment analysis extracts subjective information from text. This research evaluates machine learning (ML) techniques for understanding Arabic emotions. Sentiment analysis (SA) uses a support vector machine (SVM), AdaBoost classifier (AC), maximum entropy (ME), k-nearest neighbors (KNN), decision tree (DT), random forest (RF), logistic regression (LR), and naive Bayes (NB). A model for the ensemble-based sentiment was developed. Ensemble classifiers (ECs) with 10-fold cross-validation out-performed other machine learning classifiers in accuracy (A), specificity (S), precision (P), F1 score (FS), and sensitivity (S).

Corresponding Author:

Nouri Hicham
Research Laboratory on New Economy and Development, Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University of Casablanca
Casablanca, Morocco
Email: nhicham191@gmail.com

1. INTRODUCTION

The content of this section is divided into two main parts. The first part will be devoted to the presentation of the Arabic language. On the other hand, the second part presents the sentiment analysis in this language.

Language is the most prevalent medium via which individuals can communicate, chat, discuss, write, and express their thoughts and emotions. There are thousands of different languages in use across the globe, and it is a given that every nation has its official language. One of these languages is Arabic [1], which belongs to the Semitic language family, and Hebrew and Aramaic. Arabs speak Arabic as their first language around 260 million people, and even more, individuals comprehend Arabic as a second language [2]. The Arabic alphabet is distinct from other alphabets and is written from right to left, just like the Hebrew alphabet. Arabic is one of the six languages the United Nations has to deal with, along with English, French, Spanish, Russian, and Chinese. This is because Arabic is one of the most widely spoken languages in the entire globe. Arabic is recognized as an official language in several nations, even though it is sometimes spoken differently. This language is spoken in several different variants or dialects, including Modern Standard, Egyptian, Gulf, Moroccan, and Levantine [2].

Because of the significant differences between several dialects, they might need help comprehending one another. There are 28 letters in the Arabic alphabet, each of which can take on a variety of shapes and be spelled in various ways depending on where it is located within a word. There is a "classical" form of the Arabic language known as modern standard Arabic (MSA) [2]. The Holy Quran, other books, news, government publications, and periodicals are all written in MSA, the language used throughout the Arab world. However, when speaking with one another in day-to-day life, a slang variant of the "street language" is employed instead of the standard form. For instance, Jordan's individuals have their dialect and a very particular way of communicating verbally. Locals speak the Khaliji language in the Arabian Gulf region. Those who live in Lebanon and Syria speak a dialect known as Levantine, which is also known as (Shammi). In contrast, people who live in Morocco, Libya, Algeria, and Tunisia speak a dialect known as Maghrebi. Similarly, Sudanese people have a unique dialect [1], [3].

The purpose of conducting sentiment analysis is to learn the perspectives held by a population of individuals through one or more online platforms. In recent years, there has been a precipitous rise in the number of postings made on social media. In particular, news items written in Arabic on social media have garnered significant influence among users of those platforms. Based on this, one observes and analyzes social media users' behavior, acceptance, reaction, and engagement. In addition, the research that went into developing our proposed system makes it possible for users to feel less stressed out and for relevant organizations to become better acquainted with one another. In the latter stages of this investigation, the users' opinions are characterized by their behaviors, approval, emotions, and interactions [4].

One of the numerous subfields under natural language processing (NLP) is sentiment analysis [5]. This subfield applies probabilistic logic to the study of human language. The primary objective of sentiment analysis is to categorize the documents and characterize the polarity of their content, which might be positive, negative, or neutral.

Due to the large data set that must be obtained to receive the results, one of the most difficult aspects of sentiment analysis is the collection of the data [6]. Because of this, many application programming interfaces (API) exist. The ability to quickly and effectively acquire datasets from various sources is the primary focus of these application programming interfaces (APIs) [7]. Application programming interfaces, also known as APIs, are a collection of programs that share a common communication protocol. This protocol is responsible for the primary function of data collection, which is necessary for the development of software such as sentiment analysis or any other program that necessitates the generation of a large amount of data. APIs are a group of applications. APIs are composed of a group of programs that each have its communication mechanism.

The use of sentiment analysis has a wide range of applications [8], particularly in the commercial world. The areas of business intelligence application and recommender systems benefit the most from sentiment analysis [9]. One of these advantages is the capability to analyze the responses, comments, feedback, and contributions made by users of social media platforms to monitor the social media sites of trademark brands and companies. The intention behind sentiment analysis is to glean or anticipate the polarity of people's behaviors, emotions, approval, and relationships with one another.

Based on the ensemble machine learning technique, we have created a model for sentiment analysis in Arabic in this study using a support vector machine (SVM), AdaBoost classifier (AC), maximum entropy (ME), k-nearest neighbors (KNN), decision tree (DT), random forest (RF), logistic regression (LR), and naive Bayes (NB). Following the completion of the optimization procedure, the developed model's performance was superior to that of the other single models. This is an abbreviated version of our contributions, which can be summarized as: i) for machine learning, we suggested utilizing different algorithms such as AC, SVM, ME, DT, KNN, RF, NB, and LR; ii) this study explores unigram, bigram, and term frequency-inverse document frequency (TF-IDF) for feature extraction (FE); iii) we present a model for an ensemble machine-learning technique that incorporates several algorithms; and iv) this model is evaluated using four distinct datasets.

In the part that follows, some background information will be presented. In section 3, we shall discuss the proposed method. Section 4 will present the analytic system's evaluation findings and use case. In conclusion, the summary and future directions are discussed in the concluding section.

2. RELATED WORK

The number of models and algorithms that can conduct SA has increased rapidly in recent years due to the expansion of social networking and online shopping websites. This section of the article provides a comprehensive review of the most recent studies concerning SA. In recent research, sentiment analysis has been carried out using methods that are based on machine learning.

An approach for analyzing sentiment is used for datasets gathered from Arabic social media, in which a corpus is manually built, as detailed in [10]. The outcomes of this research are reported in [11]. Several ML techniques are executed on the training and testing datasets to ascertain whether or not the suggested approach produces accurate results. According to the information presented in [11], a mechanism for doing sentiment analysis is provided for processing sentiments produced from learning and teaching datasets. The processed data are subjected to a feature selection before the generation of four models, followed by an implementation of a SVM technique to provide the models with accurate findings. It is demonstrated in [2] how a NB classifier can be used for Arabic tweets. This classification method makes use of phrase frequency techniques. The testing datasets are divided into five distinct portions, and those divisions are then utilized to categorize the opinions expressed in the tweets. SVM are another method for evaluating sentiment [1]. These machines classify the text datasets that are provided. The statistics are then averaged after five rounds of testing to assess how well they did in the test. According to what is mentioned in [12], sentiment categorization is accomplished by using a variety of machine learning algorithms, and the effectiveness of the various machine learning approaches is assessed.

Jose and Chooralil [13] offer an automatic classifier that combines lexicon-based techniques with machine learning methods. During this research, the dataset is gathered and preprocessed, and a lexicon is built with the help of Senti WordNet. Poornima and Priya [14] classified the polarity of the emotion based on the frequency of certain phrases within the dataset. The dataset is partitioned into training and testing label columns to evaluate the effectiveness of various machine-learning strategies. According to the hypothesis presented in [15], feature extraction methodologies are specified in conjunction with approaches to SA. An analysis of feature extraction strategies is presented here, using a wide variety of dataset domains. As seen in [16], the NB and SVM techniques are utilized to classify customer behaviors inside the e-commerce dataset. The effectiveness of the machine learning algorithms is evaluated, and the results show that NB performs better than SVM. According to the findings of Mamun *et al.* [17], the accuracy of the ensemble approached (LR+SVM+RF) with frequency-inverse document frequency features was 82% higher than that of the other simple classifier models on the dataset that was constructed. Figure 1 illustrates the several machine-learning approaches for SA [8].

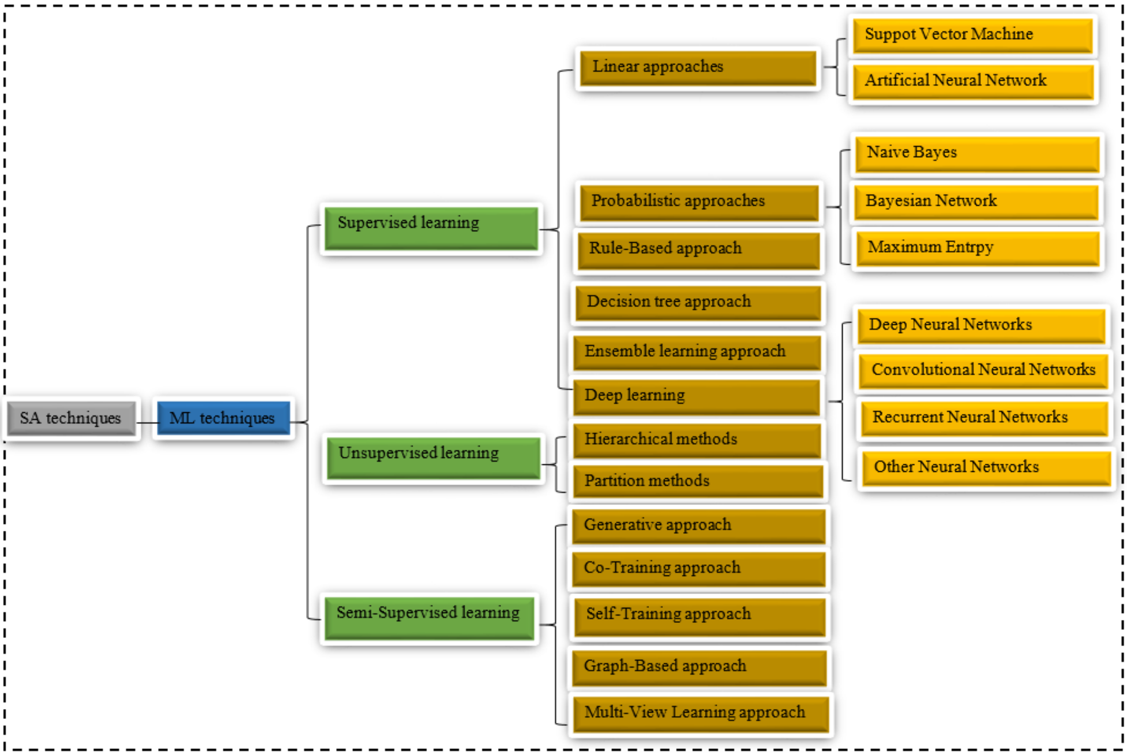

Figure 1. Sentiment analysis techniques in machine learning

3. TOOLS AND METHOD

Following these sentences, we will explore the machine learning algorithms that we have been using. Datasets, processing of the datasets, and feature extraction. But before we go into that, let us start with the core framework of our study, which is depicted in Figure 2 (a detailed description of this figure is presented in paragraphs 3.1, 3.2, 3.3, and 3.4).

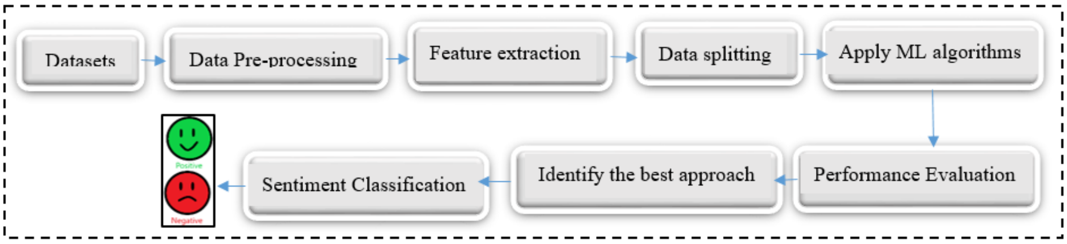

Figure 2. Conceptual framework for sentiment analysis

3.1. Dataset used

This is one of the most crucial elements in sentiment analysis. Everything will be determined by the quality of the obtained data and how it is utilized. This phase introduces the first sentiment analysis aspect. In this work, the following datasets were used to test the efficacy of our model: the trip advisor (TA) dataset [18], the booking of hotels for Arabic (BHA) dataset [19], the SentiWordNet (SWN) dataset [20], and Arabic sentiment (AraSenti) tweet [21]. The three different datasets utilized for this research are outlined in Table 1.

Table 1. Our model's evaluation datasets

	Databases			
	TA	BHA	AraSenti	SWN
Positive	1,000	4,112	5,066	15,414
Negative	1,000	4,112	5,067	15,414
Total	20,000	8,224	10,133	30,828

3.2. Data pre-processing

Typically, the data obtained from many sources, notably social media, are unstructured. The raw version of these data may possess a great deal of noise and several typographical and grammatical mistakes. Therefore, text must be cleaned and preprocessed before analysis. The preprocessing phase aims not only to obtain a more accurate analysis but also to reduce the dimensionality of the input data, as many words are superfluous and should be deleted because they do not affect the text polarity. There are numerous common tasks involved in the overall process:

- Tokenization: this phase divides the text into smaller components known as tokens.
- Stop word removal: stop words are words (such as "the", "for", and "under") that do not often contribute to the analysis and are therefore eliminated in advance.
- Part-of-speech (PoS) tagging: this process identifies several structural aspects of a text, including verbs, adjectives, nouns, and adverbs.
- Lemmatization: the lemmatization process involves breaking down a word into its parts or roots. This is comparable to stemming, except that lemmatization preserves word-related information such as PoS tags.
- Depending on the input data format, the preprocessing stage can vary. Nevertheless, some forms require further processing and cleaning procedures, such as extending abbreviations and deleting repeated characters such as the "I" in "like". As previously noted, textual data can be quite noisy; hence, a key step, feature extraction, is required to do a more accurate sentiment analysis; this will be detailed next.

3.3. Feature extraction

Feature extraction, or feature engineering, is a critical step in sentiment analysis since it can directly affect the performance of sentiment classifications [22], [23]. This work aims to extract useful information (e.g., phrases expressing emotion) that describe significant aspects of the text. However, interacting with social media posts presents additional obstacles, and more functions can be included. We explore unigram, bigram, and TF-IDF for feature extraction in this study.

3.4. Machine learning models

For learning and classification, machine learning algorithms use a variety of series [24]. The training set consists of both the classes and names of the feature vectors that are being input. A classification model was created to determine whether input values are positive or negative [25]. The extracted feature sets are utilized in the training of the classifier, which assists in determining whether or not the analysis of the data set is positive. Ensemble methods are machine learning methods that combine several base models into a single model that makes the best predictions.

3.4.1. AdaBoost classifier

AdaBoost classifier (AC) is a technique that can be applied with either textual or numeric data types. It may also be utilized with both. To obtain greater degrees of accuracy, ensembles continually segment the data and keep adding new weights to the ones they have already learned from. This ensures that any data from the first split that was erroneously assigned will be reassigned correctly during the subsequent data split. This process will be carried out multiple times until the method that proves to be the most successful in partitioning the data is identified [26].

3.4.2. Support vector machine

Support vector machine (SVM) were chosen as one of the most effective classifiers despite the small number of available datasets. This binary categorization technique with discrete target variables is also one of the most frequently employed categorization methods. By modifying the kernel model, SVM can accomplish non-linear classification despite being a linear classifier. It seeks to identify the border or hyperplane that best divides the parameters of two distinct classes in a dataset. The isolating hyperplane resides within a subspace of (N-1) dimensions, where N represents the parameters numbers [27].

3.4.3. Maximum entropy

In their respective data representations, conditional exponential classifiers will either store labeled feature sets as arras or vectors of integers. After computing the feature weights with this vector, selecting the label that is most likely to apply to the feature set is done with the help of the feature weights. It is possible to think of entropy as a gauge of how random something is. The entropy will be at its highest level when the data are dispersed uniformly throughout the space. The sentences and ratings on a scale ranging 1 through 5 include the data input [28].

3.4.4. Decision tree

Decision tree (DT) is a self-education method that utilizes mathematics and builds data forecasting into a tree-like framework. This framework is acquired through the process of learning under supervision, and it is utilized in the process of deciding which choice is the most advantageous. This makes it possible to perform automatic clustering by employing a training data and data category prediction and for data that has not been assigned to a certain categories in the past. The value assigned to each attribute by the rank widget in [29] is decided by the attribute’s link to the class.

3.4.5. K-nearest neighbors

K-nearest neighbors (KNN) is used in machine learning. KNN is a supervised learning algorithm; hence, there is no training step. This algorithm classifies data by computing the distance between test and training data and determining its closest neighbors. The shortest distance between two places is calculated using Euclidean distance and cosine similarity. First method assesses Euclidean distance between training and test data; second employs cosine similarity [30].

3.4.6. Random forest

This technique creates many decision trees by employing random data sampling for learning. The results of the majority vote are then used to piece together the outputs of the decision trees. The implementation of classification, regression, and data clustering can all be accomplished through the use of a process known as ensemble learning. The model’s training is carried out in a manner that is consistent by using the random forest’s provided subset of features and the random forest’s provided subset of data. Because of this technique, the garment is less likely to become oversized. In recent years, this algorithm has seen broad use in various applications, and it has shown higher performance when compared to that another algorithm. This is due to the fact that its versatility has allowed it to be used in a wide range of contexts [31].

3.4.7. Naive Bayes

The naive Bayes (NB) classifier, one of the most well-known supervised classifiers, enables you to describe neutral, negative, or positive emotions in the text you generate. It classifies words into their respective groups based on conditional probabilities. Using NB classifier for text categorization has the time-saving advantage of requiring only a minimal dataset for training purposes. When raw data is pre-processed, among other things, stop words, punctuation marks, extra spaces, transliteration of words from other languages, and special symbols are removed. The difficult chore of manually labeling words as positive, negative, or neutral falls to human annotators [12].

3.4.8. Logistic regression

Logistic regression (LR) is a method of statistical analysis that is utilized in the process of investigating the connection that exists between a categorical dependent variable and one or more independent variables. A logistic function, known as the cumulative logistic distribution, is utilized in calculating probabilities using this methodology. Logistic regression is linear; however, the logistic function changes the predictions to produce more accurate findings. It is a statistical analysis method utilized in evaluating datasets containing one or more independent variables that affect the outcomes [32].

3.4.9. Ensemble classifier

Ensemble approaches aim to blend the predictions of many base estimators with a specific learning algorithm to improve the classifier's precision and robustness. The voting classifier's objective is to merge conceptually distinct machine learning models and forecast class labels depending on the majority of votes or the expected mean probability. Such a classifier can effectively compensate for a group of classifiers with comparable performance shortcomings [33]. A method for classifying sentiments based on ensembles and using machine learning algorithms is illustrated in Figure 2. And algorithms are AC, SVM, ME, DT, KNN, RF, NB, and LR.

4. EXPERIMENTS AND FINDINGS

In this section's first part, we will discuss the evaluation methods we adopted to evaluate our design's efficiency. Accuracy, sensitivity, specificity, and precision are the five metrics that make up this evaluation. After that, we shall proceed to the conclusions that we have drawn.

4.1. Performance measures

Our proposed model's performance and effectiveness are evaluated using different metrics. The selection of evaluation metrics can influence the manner in which the performance and efficacy of a model are compared and monitored. In our study, we used five different evaluation metrics, which are presented in Table 2 [1], [4].

Table 2. Description of evaluation metrics for sentiment analysis

Performance Measure	Description	Calculation
Accuracy	Accuracy is the ratio of accurately predicted examples to the total number of examples. This metric's counterpart, error, equals 1 accuracy.	$Acc = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$
Precision	Precision is the proportion of accurately anticipated positive samples to the total number. Precision measures precision.	$Pre = \frac{Tp}{Tp + Fp}$
F1-score	F1-score measures precision and memory by calculating their harmonic mean.	$\frac{2 * (\text{Precision.Recall})}{(\text{Precision} + \text{Reacall})}$
Recall (Sensitivity)	Recall is the percentage of positive samples properly anticipated. Measures model misclassification.	$\frac{Tp}{Tp + Fn}$

4.2. Results and discussions

In order to demonstrate the performance of our model, we tested it on the datasets mentioned earlier and compared its findings to those of other individual algorithms such as SVM, AC, ME, DT, KNN, RF, NB, and LR. For performance evaluation, we utilized the following metrics: accuracy, specificity, precision, F1-score and Sensitivity. The findings that were obtained are detailed in the paragraphs that follow.

On dataset TA as shown in Table 3, for each classifier, we performed 5-fold cross validation (cv); AC, SVM, ME, DT, KNN, RF, NB, and LR, we obtained accuracy as 0.812, 0.862, 0.756, 0.817, 0.867, 0.668, 0.671, and 0.724, respectively and obtained higher results in terms of accuracy for our model, EC as 0.871. For 10-fold cross validation, we received for every method; SVM, AC, DT, ME, KNN, RF, NB, and LR, we obtained accuracy as 0.835, 0.885, 0.779, 0.840, 0.890, 0.691, 0.694, and 0.747, respectively, and obtained higher results in terms of model accuracy, EC as 0.894.

Table 4 shows that on dataset TA, we conducted Unigram for each classifier; AC, SVM, ME, DT, KNN, RF, NB, and LR; we obtained accuracy as 0.786, 0.836, 0.730, 0.791, 0.841, 0.642, 0.645, and 0.698, respectively and obtained higher results in terms of model accuracy, EC as 0.845. For Bigram, we obtained for each classifier; AC, SVM, ME, DT, KNN, RF, NB, and LR, we obtained accuracy as 0.709, 0.759, 0.653, 0.714, 0.764, 0.565, 0.568, and 0.621 respectively and obtained higher results in terms of model accuracy, EC as 0.768. On the other hand, our model based on the novel ensemble machine learning presents the best results in the different evaluation metrics.

Table 3. Model’s performance comparison on dataset TA with k-fold validation

Classifier	TF-IDF (k=5)					TF-IDF (k=10)				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.812	0.765	0.802	0.725	0.698	0.835	0.778	0.829	0.738	0.727
SVM	0.862	0.771	0.769	0.731	0.625	0.885	0.784	0.796	0.744	0.654
ME	0.756	0.719	0.754	0.697	0.612	0.779	0.732	0.781	0.710	0.641
DT	0.817	0.723	0.801	0.765	0.697	0.840	0.736	0.828	0.778	0.726
KNN	0.867	0.827	0.829	0.781	0.701	0.890	0.840	0.856	0.794	0.730
RF	0.668	0.687	0.774	0.691	0.625	0.691	0.700	0.801	0.704	0.654
NB	0.671	0.702	0.795	0.659	0.595	0.694	0.715	0.822	0.672	0.624
LR	0.724	0.711	0.824	0.745	0.687	0.747	0.724	0.851	0.758	0.716
EC	0.871	0.867	0.873	0.824	0.879	0.894	0.880	0.900	0.837	0.908

Table 4. Model’s performance comparison on dataset TA with unigram and bigram

Classifier	Unigram					Bigram				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.786	0.740	0.765	0.694	0.669	0.709	0.663	0.688	0.617	0.592
SVM	0.836	0.746	0.732	0.700	0.596	0.759	0.669	0.655	0.623	0.519
ME	0.730	0.694	0.717	0.666	0.583	0.653	0.617	0.640	0.589	0.506
DT	0.791	0.698	0.764	0.734	0.668	0.714	0.621	0.687	0.657	0.591
KNN	0.841	0.802	0.792	0.750	0.672	0.764	0.725	0.715	0.673	0.595
RF	0.642	0.662	0.737	0.660	0.596	0.565	0.585	0.660	0.583	0.519
NB	0.645	0.677	0.758	0.628	0.566	0.568	0.600	0.681	0.551	0.489
LR	0.698	0.686	0.787	0.714	0.658	0.621	0.609	0.710	0.637	0.581
EC	0.845	0.842	0.836	0.793	0.850	0.768	0.765	0.759	0.716	0.773

On the dataset BHA as depicted in Table 5, we did 5-fold cross validation (cv) for each method; for SVM, AC, DT, ME, KNN, RF, NB, and LR, we achieved accuracy values of 0.820, 0.879, 0.773, 0.834, 0.885, 0.685, 0.688, 0.741, and 0.888 for our model, the ensemble classifier (EC). For 10-fold cross-validation, we achieved accuracy values of 0.852, 0.902, 0.796, 0.857, 0.907, 0.708, 0.711, and 0.764 for each classifier; AC, SVM, ME, DT, KNN, RF, NB, and LR; and produced superior results in terms of accuracy for our model, EC, which was 0.911%. On the other hand, our model’s results, based on an innovative form of machine learning ensemble, are the best across all of the different evaluation metrics.

Table 5. Model’s performance comparison on dataset BHA with k-fold validation

Classifier	TF-IDF (k=5)					TF-IDF (k=10)				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.829	0.782	0.819	0.742	0.715	0.852	0.795	0.846	0.755	0.744
SVM	0.879	0.788	0.786	0.748	0.642	0.902	0.801	0.813	0.761	0.671
ME	0.773	0.736	0.771	0.714	0.629	0.796	0.749	0.798	0.727	0.658
DT	0.834	0.740	0.818	0.782	0.714	0.857	0.753	0.845	0.795	0.743
KNN	0.884	0.844	0.846	0.798	0.718	0.907	0.857	0.873	0.811	0.747
RF	0.685	0.704	0.791	0.708	0.642	0.708	0.717	0.818	0.721	0.671
NB	0.688	0.719	0.812	0.676	0.612	0.711	0.732	0.839	0.689	0,641
LR	0.741	0.728	0.841	0.762	0.704	0.764	0.741	0.868	0.775	0.733
EC	0.888	0.884	0.890	0.841	0.896	0.911	0.897	0.917	0.854	0.925

On the BHA dataset, as shown in Table 6, we did unigram for each classifier; AC, SVM, ME, DT, KNN, RF, NB, and LR; we achieved accuracy values of 0.827, 0.877, 0.771, 0.832, 0.882, 0.683, 0.686, and 0.739; and produced superior results in terms of accuracy for our model’s EC, which was 0.886. For bigram, we achieved the following results for each classifier: AC, SVM, ME, DT, KNN, NB, RF, and LR: 0.750, 0.800, 0.694, 0.755, 0.805, 0.606, 0.609, 0.662 and received superior results in terms of accuracy for our model’s EC: 0.809%. On the contrary, our model’s results, based on an innovative form of machine learning ensemble, are the best across all of the various performance metrics.

As seen in Table 7, for the dataset AraSenti, we performed 5-fold cross validation (cv) for each technique; for SVM, AC, DT, ME, KNN, RF, NB, and LR, we obtained accuracy values of 0.840, 0.890, 0.784, 0.845, 0.895, 0.696, 0.699, 0.752 and obtained superior results in terms of accuracy for our model, EC as 0.899%. For 10-fold cross-validation, we achieved accuracy values of 0.863, 0.913, 0.807, 0.868, 0.918, 0.719, 0.722, and 0.775 for each classifier; SVM, AC, DT, ME, KNN, RF, NB, and LR; and produced superior results in terms of accuracy for our model, ensemble classifier as 0.922%. In contrast, our innovative ensemble machine learning model achieves top performance across all measures we tested.

Table 6. Model's performance comparison on dataset BHA with unigram and bigram

Classifier	Unigram					Bigram				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.827	0.781	0.806	0.735	0.710	0.750	0.704	0.729	0.658	0.633
SVM	0.877	0.787	0.773	0.741	0.637	0.800	0.710	0.696	0.664	0.560
ME	0.771	0.735	0.758	0.707	0.624	0.694	0.658	0.681	0.630	0.547
DT	0.832	0.739	0.805	0.775	0.709	0.755	0.662	0.728	0.698	0.632
KNN	0.882	0.843	0.833	0.791	0.713	0.805	0.766	0.756	0.714	0.636
RF	0.683	0.703	0.778	0.701	0.637	0.606	0.626	0.701	0.624	0.560
NB	0.686	0.718	0.799	0.669	0.607	0.609	0.641	0.722	0.592	0.530
LR	0.739	0.727	0.828	0.755	0.699	0.662	0.650	0.751	0.678	0.622
EC	0.886	0.883	0.877	0.834	0.891	0.809	0.806	0.800	0.757	0.814

Table 7. Model's performance comparison on dataset AraSenti with k-fold validation

Classifier	TF-IDF (k=5)					TF-IDF (k=10)				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.840	0.793	0.830	0.753	0.726	0.863	0.806	0.857	0.766	0.755
SVM	0.890	0.799	0.797	0.759	0.653	0.913	0.812	0.824	0.772	0.682
ME	0.784	0.747	0.782	0.725	0.640	0.807	0.760	0.809	0.738	0.669
DT	0.845	0.751	0.829	0.793	0.725	0.868	0.764	0.856	0.806	0.754
KNN	0.895	0.855	0.857	0.809	0.729	0.918	0.868	0.884	0.822	0.758
RF	0.696	0.715	0.802	0.719	0.653	0.719	0.728	0.829	0.732	0.682
NB	0.699	0.730	0.823	0.687	0.623	0.722	0.743	0.850	0.700	0.652
LR	0.752	0.739	0.852	0.773	0.715	0.775	0.752	0.879	0.786	0.744
EC	0.899	0.895	0.901	0.852	0.907	0.922	0.908	0.928	0.865	0.936

Table 8 presents the findings based on the dataset AraSenti. We did unigram for each classifier; AC, SVM, ME, DT, KNN, RF, NB, and LR, obtaining accuracy values of 0.834, 0.884, 0.778, 0.839, 0.889, 0.690, 0.693, and 0.746 correspondingly and achieving superior results in terms of accuracy for our model, EC as 0.893. For bigram, we received for each classifier; AC, SVM, ME, DT, KNN, RF, NB, and LR, accuracy values of 0.757, 0.807, 0.701, 0.762, 0.812, 0.613, 0.616, and 0.669, respectively, and produced superior results in terms of accuracy for our model's EC, which was 0.816%. Our model based on innovative ensemble machine learning achieves the highest scores across all evaluation metrics.

According to the SWN dataset displayed in Table 9, we performed 5-fold cross validation (cv) for each technique; for SVM, AC, DT, ME, KNN, RF, NB, and LR, we obtained accuracy values of 0.855, 0.905, 0.799, 0.860, 0.910, 0.711, 0.714, and 0.767 respectively, with our ensemble classifier (EC) achieving superior results in terms of accuracy at 0.914%. For 10-fold cross-validation, we acquired each classifier's accuracy values: SVM, AC, DT, ME, KNN, RF, NB, and LR: 0.878, 0.928, 0.822, 0.883, 0.933, 0.734, 0.737, and 0.790. Our model's ensemble classifier achieved superior results in terms of accuracy, achieving 0.937. On the other hand, our model based on innovative ensemble machine learning achieves the greatest performance across the various evaluation measures.

According to the SWN dataset, which is outlined in Table 10, we ran the unigram algorithm for each of the following classifiers: AC, SVM, ME, DT, KNN, RF, NB, and LR. The resulting accuracy values for these classifiers were as: 0.847, 0.897, 0.791, 0.852, 0.902, 0.703, 0.706, and 0.759, respectively. However, the results obtained by our model's ensemble classifier were significantly better, coming in at 0.906. For Bigram, we found that the accuracy of each classifier; AC, SVM, ME, DT, KNN, RF, NB, and LR-ranged from 0.770, 0.820, 0.714, 0.775, 0.825, 0.626, 0.629, and 0.682, respectively. However, we found that the accuracy of our model's ensemble classifier, which was 0.829, was superior to any of the individual classifiers. Our model, on the other hand, based on the novel ensemble machine learning, achieves the greatest performance across all evaluation metrics.

Table 8. Model's performance comparison on dataset AraSenti with unigram and bigram

Classifier	Unigram					Bigram				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.834	0.788	0.813	0.742	0.717	0.757	0.711	0.736	0.665	0.640
SVM	0.884	0.794	0.780	0.748	0.644	0.807	0.717	0.703	0.671	0.567
ME	0.778	0.742	0.765	0.714	0.631	0.701	0.665	0.688	0.637	0.554
DT	0.839	0.746	0.812	0.782	0.716	0.762	0.669	0.735	0.705	0.639
KNN	0.889	0.850	0.840	0.798	0.720	0.812	0.773	0.763	0.721	0.643
RF	0.690	0.710	0.785	0.708	0.644	0.613	0.633	0.708	0.631	0.567
NB	0.693	0.725	0.806	0.676	0.614	0.616	0.648	0.729	0.599	0.537
LR	0.746	0.734	0.835	0.762	0.706	0.669	0.657	0.758	0.685	0.629
EC	0.893	0.890	0.884	0.841	0.898	0.816	0.813	0.807	0.764	0.821

Table 9. Model's performance comparison on dataset SWN with k-fold validation

Classifier	TF-IDF (k=5)					TF-IDF (k=10)				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.855	0.808	0.845	0.768	0.741	0.878	0.821	0.872	0.781	0.770
SVM	0.905	0.814	0.812	0.774	0.668	0.928	0.827	0.839	0.787	0.697
ME	0.799	0.762	0.797	0.740	0.655	0.822	0.775	0.824	0.753	0.684
DT	0.860	0.766	0.844	0.808	0.740	0.883	0.779	0.871	0.821	0.769
KNN	0.910	0.870	0.872	0.824	0.744	0.933	0.883	0.899	0.837	0.773
RF	0.711	0.730	0.817	0.734	0.668	0.734	0.743	0.844	0.747	0.697
NB	0.714	0.745	0.838	0.702	0.638	0.737	0.758	0.865	0.715	0.667
LR	0.767	0.754	0.867	0.788	0.730	0.790	0.767	0.894	0.801	0.759
EC	0.914	0.910	0.916	0.867	0.922	0.937	0.923	0.943	0.880	0.951

Table 10. Model's performance comparison on dataset SWN with unigram and bigram

Classifier	Unigram					Bigram				
	A	S	P	FS	S	A	S	P	FS	S
AC	0.847	0.801	0.826	0.755	0.730	0.770	0.724	0.749	0.678	0.653
SVM	0.897	0.807	0.793	0.761	0.657	0.820	0.730	0.716	0.684	0.580
ME	0.791	0.755	0.778	0.727	0.644	0.714	0.678	0.701	0.650	0.567
DT	0.852	0.759	0.825	0.795	0.729	0.775	0.682	0.748	0.718	0.652
KNN	0.902	0.863	0.853	0.811	0.733	0.825	0.786	0.776	0.734	0.656
RF	0.703	0.723	0.798	0.721	0.657	0.626	0.646	0.721	0.644	0.580
NB	0.706	0.738	0.819	0.689	0.627	0.629	0.661	0.742	0.612	0.550
LR	0.759	0.747	0.848	0.775	0.719	0.682	0.670	0.771	0.698	0.642
EC	0.906	0.903	0.897	0.854	0.911	0.829	0.826	0.820	0.777	0.834

The currently available experimental data make it possible to investigate the connection between the success of a specific classifier and the validity of our approach. These findings demonstrate how well each technique performs on the five criteria metrics. In this part of the article, a thorough investigation was carried out to evaluate the efficiency of the proposed method. In the datasets, SWN, AraSenti, dataset BHA, and TA database, our model with the parameters Unigram, Bigram, k=5, and k=10 produces fascinating results, as can be shown in Figures 3, 4, 5 and 6. These results are based on evaluation metrics.

Our model, based on a machine learning ensemble, produces extremely significant results at the level of unigrams, bigrams, k=5, and k=10; nevertheless, it is the k=10 level that produces the best results. This model is evaluated using four distinct datasets. Therefore, to summarize, we present a model for an ensemble machine-learning technique that incorporates several algorithms. For machine learning, we suggested utilizing different algorithms such as AC, SVM, ME, DT, KNN, RF, NB, and LR. This study explores unigram, bigram, and TF-IDF for feature extraction (FE).

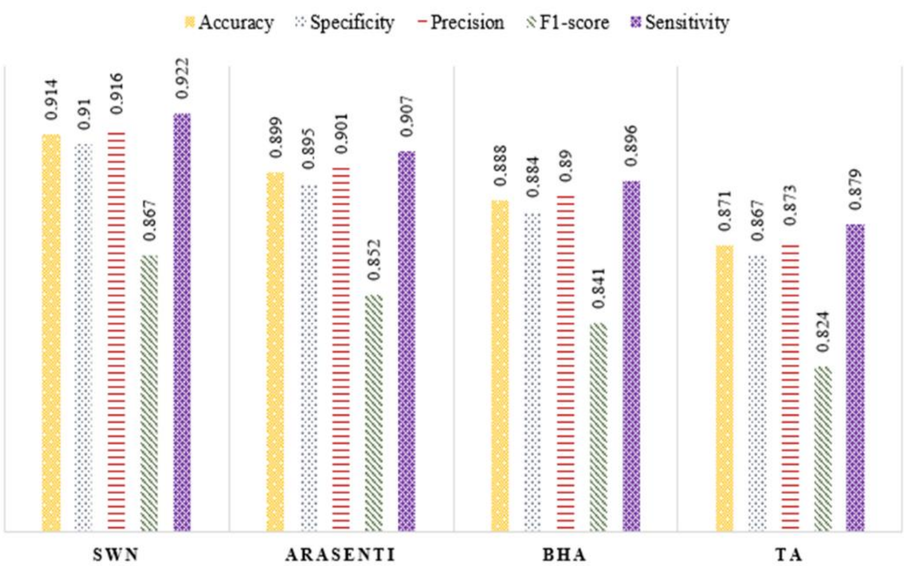

Figure 3. Performance of our model with unigram

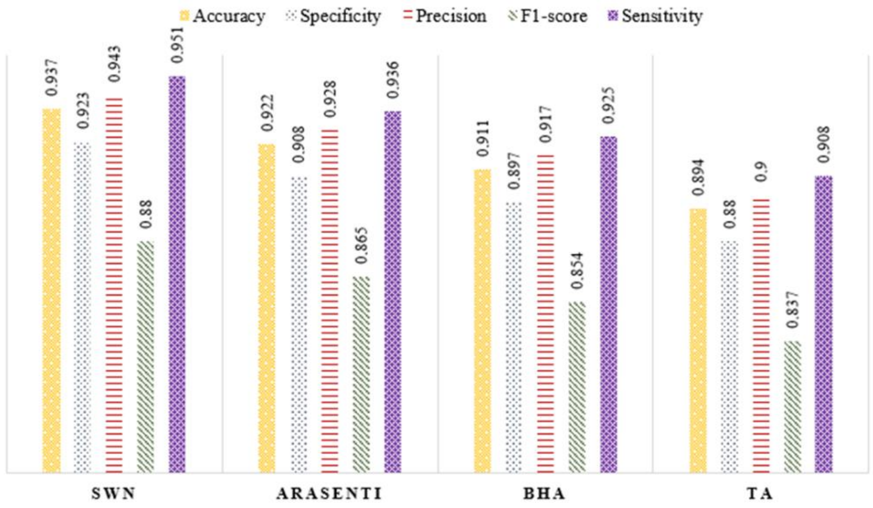

Figure 4. Performance of our model with bigram

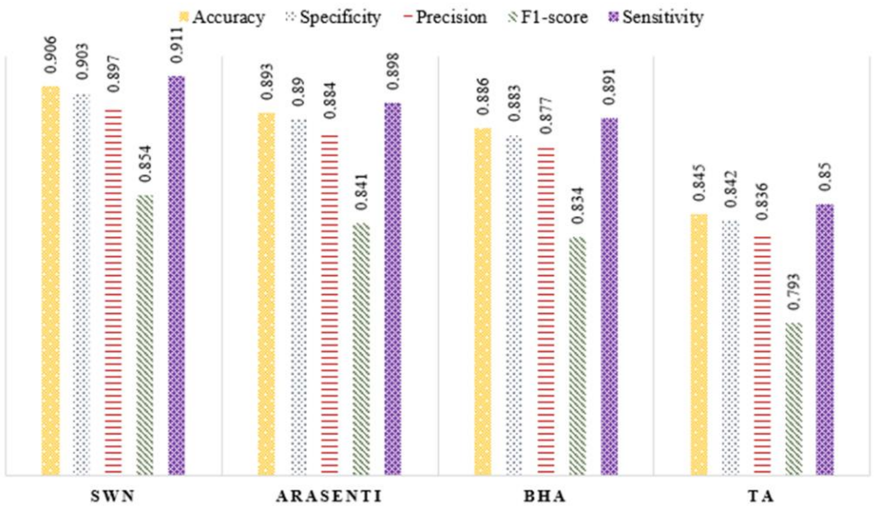

Figure 5. Performance of our model with *k=5*

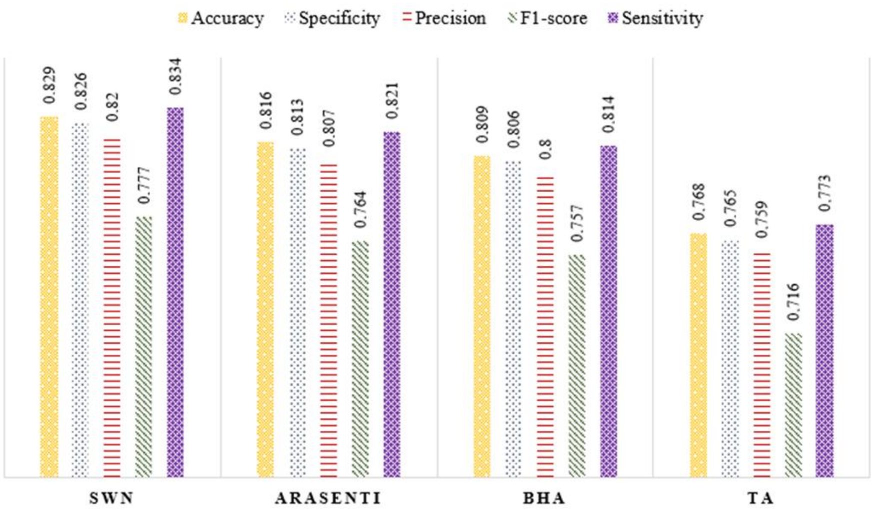

Figure 6. Performance of our model with *k=10*

5. CONCLUSION

Nowadays, Arabic sentiment analysis (ASA) is vital in various industries, including politics, services, and production. Social networks are rich with Arabic writings in which individuals communicate their viewpoints on various topics of great interest to ASA academics. In the ASA domain, machine learning techniques are crucial for data processing. In this situation, eight different classifier techniques were used: SVM, AC, ME, DT, KNN, RF, NB, and LR. Our ensemble machine learning techniques are compared to these models to identify a group of comments and positive or negative reviews.

This research began with the preliminary examination of four database files. Accuracy, Specificity, Precision, F1-score, and sensitivity were employed to evaluate the performance of the classifiers applied in this study. A 10-fold cross-validation ensemble classifier demonstrated higher performance across all criteria during ensemble-based Sentiment categorization testing. The AraBERT model will be applied to a large collection of reviews and comments written in Arabic language and various Arabic dialects in future studies. In this instance, we will examine how effectively one of the most advanced models, AraBERT, improves ASA.

REFERENCES

[1] N. Habbat, H. Anoun, and L. Hassouni, "Sentiment analysis and topic modeling on Arabic twitter data during covid-19 pandemic," *Indonesian Journal of Innovation and Applied Sciences (IJIAS)*, vol. 2, no. 1, pp. 60–67, Feb. 2022, doi: 10.47540/ijias.v2i1.432.

[2] G. Alwakid, T. Osman, M. El Haj, S. Alanazi, M. Humayun, and N. U. Sama, "MULDASA: multifactor lexical sentiment analysis of social-media content in nonstandard Arabic social media," *Applied Sciences*, vol. 12, no. 8, Apr. 2022, doi: 10.3390/app12083806.

[3] T. Potrawa and A. Tetereva, "How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market," *Journal of Business Research*, vol. 144, pp. 50–65, May 2022, doi: 10.1016/j.jbusres.2022.01.027.

[4] 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 (IJEEI)*, vol. 13, no. 4, pp. 801–812, Dec. 2021, doi: 10.15676/ijeei.2021.13.4.3.

[5] T. H. Alwaneen, A. M. Azmi, H. A. Aboalsamh, E. Cambria, and A. Hussain, "Arabic question answering system: a survey," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 207–253, Jan. 2022, doi: 10.1007/s10462-021-10031-1.

[6] S.-O. Proksch, W. Lowe, J. Wäckerle, and S. Soroka, "Multilingual sentiment analysis: a new approach to measuring conflict in legislative speeches," *Legislative Studies Quarterly*, vol. 44, no. 1, pp. 97–131, Feb. 2019, doi: 10.1111/lsq.12218.

[7] N. O. Alsrehin, A. F. Klaib, and A. Magableh, "Intelligent transportation and control systems using data mining and machine learning techniques: a comprehensive study," *IEEE Access*, vol. 7, pp. 49830–49857, 2019, doi: 10.1109/ACCESS.2019.2909114.

[8] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, Nov. 2020, doi: 10.1016/j.future.2020.05.034.

[9] F. Abbasi and A. Khadivar, "Collaborative filtering recommendation system through sentiment analysis," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 14, pp. 1843–1853, 2021.

[10] A. A. Hnaif, E. Kanan, and T. Kanan, "Sentiment analysis for Arabic social media news polarity," *Intelligent Automation and Soft Computing*, vol. 28, no. 1, pp. 107–119, 2021, doi: 10.32604/iasc.2021.015939.

[11] A. Muhammad, S. Abdullah, and N. Samsiah Sani, "Optimization of sentiment analysis using teaching-learning based algorithm," *Computers, Materials and Continua*, vol. 69, no. 2, pp. 1783–1799, 2021, doi: 10.32604/cmc.2021.018593.

[12] M. Wongkar and A. Angdresey, "Sentiment analysis using naive Bayes algorithm of the data crawler: twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–5, doi: 10.1109/ICIC47613.2019.8985884.

[13] R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, Mar. 2016, pp. 64–67, doi: 10.1109/SAPIENCE.2016.7684133.

[14] A. Poornima and K. S. Priya, "A comparative sentiment analysis of sentence embedding using machine learning techniques," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2020, pp. 493–496, doi: 10.1109/ICACCS48705.2020.9074312.

[15] M. W.Habib and Z. N. Sultani, "Twitter sentiment analysis using different machine learning and feature extraction techniques," *Al-Nahrain Journal of Science*, vol. 24, no. 3, pp. 50–54, Sep. 2021, doi: 10.22401/ANJS.24.3.08.

[16] M. Mamatha, J. Thriveni, and K. R. Venugopal, "Techniques of sentiment classification, emotion detection, feature extraction and sentiment analysis a comprehensive review," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 1, pp. 244–261, Jan. 2018, doi: 10.26438/ijcse/v6i1.244261.

[17] M. M. R. Mamun, O. Sharif, and M. M. Hoque, "Classification of textual sentiment using ensemble technique," *SN Computer Science*, vol. 3, no. 1, Jan. 2022, doi: 10.1007/s42979-021-00922-z.

[18] A. El-Halees and A. Al-Asmar, "Ontology based Arabic opinion mining," *Journal of Information and Knowledge Management*, vol. 16, no. 3, Sep. 2017, doi: 10.1142/S0219649217500289.

[19] W. A. M. Ahmed and A. M. El-Halees, "Arabic opinion mining using parallel decision trees," in *2017 Palestinian International Conference on Information and Communication Technology (PICICT)*, May 2017, pp. 46–52, doi: 10.1109/PICICT.2017.28.

[20] F. H. H. Mahyoub, M. A. Siddiqui, and M. Y. Dahab, "Building an Arabic sentiment lexicon using semi-supervised learning," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 417–424, Dec. 2014, doi: 10.1016/j.jksuci.2014.06.003.

[21] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis-a hybrid scheme," *Journal of Information Science*, vol. 42, no. 6, pp. 782–797, Dec. 2016, doi: 10.1177/0165551515610513.

[22] M. Kasri, M. Birjali, and A. Beni-Hssane, "A comparison of features extraction methods for Arabic sentiment analysis," in *Proceedings of the 4th International Conference on Big Data and Internet of Things*, Oct. 2019, pp. 1–6, doi: 10.1145/3372938.3372998.

[23] M. Avinash and E. Sivasankar, "A study of feature extraction techniques for sentiment analysis," in *Advances in Intelligent Systems and Computing*, Springer Singapore, 2019, pp. 475–486.

[24] 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, 2022, doi: 10.14569/IJACSA.2022.0131016.

[25] R. Vincent, P. Bhatia, M. Rajesh, A. K. Sivaraman, and M. Al Bahri, "Indian currency recognition and verification using transfer learning," *International Journal of Mathematics and Computer Science*, vol. 15, no. 4, pp. 1279–1284, 2020.

[26] M. Li, P. Xiao, and J. Zhang, "Text classification based on ensemble extreme learning machine," *arXiv preprint arXiv:1805.06525*, 2018.

[27] N. N. W. Nik Hashim, N. A. Basri, M. A.-E. Ahmad Ezzi, and N. M. H. Nik Hashim, "Comparison of classifiers using robust features for depression detection on Bahasa Malaysia speech," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 238–253, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp238-253.

[28] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[29] S. Tangwannawit and P. Tangwannawit, "An optimization clustering and classification based on artificial intelligence approach for internet of things in agriculture," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 201–209, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp201-209.

[30] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[31] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved random forest for classification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012–4024, Aug. 2018, doi: 10.1109/TIP.2018.2834830.

[32] X. Hu, Y. Yang, S. Zhu, and L. Chen, "Research on a hybrid prediction model for purchase behavior based on logistic regression and support vector machine," in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, May 2020, pp. 200–204, doi: 10.1109/ICAIBD49809.2020.9137484.

[33] 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)*, Dec. 2022, pp. 1–6, doi: 10.1109/CommNet56067.2022.9993924.

BIOGRAPHIES OF AUTHORS

Nouri Hicham is a Moroccan doctoral student at Hassan II University. In 2021, he received a master's degree in spatial economy and territorial governance from the Faculty of Legal Economic and Social Sciences at AIN SEBAA, as well as a master's degree in data engineering from the Hassania School of Public Works. My fields of study are artificial intelligence and marketing. He can be contacted at email: nourihicham@ieee.org.

Sabri Karim Doctor of Economics and Management from HEC-Ulg, the University of Liège's Management School. He has been an economics professor at the French Community in Belgium since 2003. Since 2011, he has been an economics and management professor in the Faculty of Legal, Economic, and Social Sciences at FSJES-AS Casablanca-Ain Sebaa. He is currently interested in the field of marketing, performance and IA. He can be contacted at email: sabrikarimprof@gmail.com.

Nassera Habbat Ph.D. student at Hassan II University in Morocco. In 2015, she got his Master's in Information Systems Engineering from Caddi Ayyad University. In 2013, she got his Technology and Web Programming license from the same school. Big data and machine learning are two things she is interested in studying. She can be contacted at email: nassera.habbat-etu@etu.univh2c.ma.