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

Revue d'Intelligence Artificielle
Vol. 38, No. 6, December, 2024, pp. 1399-1407
Journal homepage: http://iieta.org/journals/ria

A Suitable Technique for Enhancing Arabic-Language Consumer Sentiment Analysis Using Natural Language Processing and Stacking Machine Learning Model

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

[1] Faculty of Legal Economic and Social Sciences AIN SEBAA, Research Laboratory on New Economy and Development (LARNED), Hassan II University of Casablanca, Casablanca 2634, Morocco
[2] Faculty of Science and Technology of Settat, Hassan First University, Settat 577, Morocco

Corresponding Author Email: nhicham191@gmail.com

https://doi.org/10.18280/ria.380602

Received: 18 August 2023
Revised: 12 June 2024
Accepted: 23 July 2024
Available online: 21 December 2024

Keywords:
Arabic language, sentiment analysis, NLP, machine learning, stacked model

ABSTRACT

When deciding on a product, sentiments expressed on social media or online reviews are important information sources. Positive and negative feedback from customers posted on social media platforms could substantially impact a business's bottom line. As a result, the development of effective and efficient approaches for classifying emotion has emerged as one of the most pressing concerns for businesses. Applying machine learning is widely regarded as one of the most effective and beneficial ways. This work will investigate how well Machine Learning (ML) techniques can comprehend Arabic sentiments. The Term Frequency-Inverse Document Frequency algorithm (TF-IDF) was used to extract the dataset's characteristics. As a consequence of this, the algorithms known as Random Forest (RF), Decision Tree (DT), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), support vector machine (SVM), Quadratic Discriminant Analysis (QDA), logistic regression (LR), Gradient Boosting Regression Trees (GBRT), and Stochastic Gradient Descent (SGD) Classifier are used in the process of sentiment analysis (SA). To sum everything up, a stacked machine-learning model was developed. Compared to existing machine learning simple classifiers, our stacked model with 10-fold cross-validation shows a higher accuracy, precision, Cohen's Kappa, recall, and F1-score in the three different Arabic datasets used, which are the Hotel Arabic-Reviews Dataset (A), the Books Reviews in Arabic Dataset (B), and the Arabic Reviews dataset (C).

1. INTRODUCTION

Sentiment analysis in Arabic language involves using natural language processing (NLP) techniques to analyze and classify the emotional tone of text written in Arabic [1]. The goal is to identify the sentiment of the text as negative, neutral, or positive. There are several challenges to performing sentiment analysis in Arabic, including the complexity of the language, the lack of standardized spelling and grammar rules, and the presence of dialects and regional variations. However, several approaches can be used to address these challenges, including machine learning techniques, rule-based systems, and lexicon-based methods [2].

Consumer sentiment analysis (CSA) is a vital aspect of business intelligence, marketing, and customer experience management. It involves the extraction of opinions, attitudes, and emotions expressed by consumers in various forms of communication, such as social media, reviews, and customer support interactions [3]. With the growth of e-commerce and social media, companies are increasingly turning to sentiment analysis to gain insights into their customers' behavior and preferences. However, analyzing large volumes of unstructured text data is a complex and time-consuming task that requires sophisticated machine learning (ML) approaches and NLP [4]. CSA is a vital instrument for companies to use to comprehend their customers' thoughts and opinions regarding the services or products that the company provides. By analyzing customer feedback, businesses can gain insights into areas of improvement, identify customer pain points, and make informed decisions to improve customer experience. This, in turn, can lead to increased customer loyalty, customer retention, and improved bottom lines [5].

One typical way to sentiment analysis in Arabic is to utilize ML techniques such as neural networks (NN) or SVM to divide the text into different sentiment categories. These algorithms are trained on labeled datasets of Arabic text to learn the features and patterns that are associated with different sentiment categories [6].

Recent research [3, 4] demonstrated that the application of sentiment analysis has grown to incorporate text and visual data. This problem is related to the recognition of emotions within the field of research known as affective computing [7]. Affective computing and sentiment analysis are two areas that are extremely important to Artificial Intelligence development [8] technology and have a great deal of untapped potential in a variety of different fields. Classification issues arise when ascertaining whether a piece of writing conveys an excellent or negative mood [9]. Yet, sentiment analysis is a complex procedure; instead, it requires analyzing various natural language processing (NLP) subtasks, such as recognizing subjectivity and sarcasm [10]. In addition, the writing might not have a structure, and it has the potential to have errors and colloquialisms [11].

Researchers, businesses, governments, and other organizations have all acknowledged the significance of sentiment analysis [1]. Online resources such as blogs, message boards, social networks, and wikis have emerged as significant sources of information in recent years as the number of people who use the internet has continued to rise. Because the perspectives and viewpoints presented in these online resources are highly pertinent to our day-to-day lives, it is vital to analyze this data utilizing automated public opinion monitoring to facilitate decision-making [12]. For instance, posts made on Twitter have been studied to ascertain the results of elections [13].

As a direct consequence of this, the topic of sentiment analysis has received a considerable amount of interest within the scientific world over the previous fifteen years. Since 2004, the discipline of SA has established itself with the greatest rate of expansion and level of activity. In the past few years, there has been a discernible spike in articles focusing on opinion mining and sentiment analysis [13], which proves this is the case. Google Trends indicates that there has been a rise in the number of people interested in sentiment analysis in recent years.

This research aims to analyze the efficiency of a variety of ML approaches in terms of interpreting the sentiments communicated in Arabic. In addition, the TF-IDF technique was utilized to extract features from the dataset. As a direct consequence of this, the techniques of RF, DT, LDA, KNN, SVM, QDA, LR, GBRT, and SGD Classifier were utilized in the process of sentiment analysis (SA) in the three different Arabic datasets used, which are the Hotel Arabic-Reviews Dataset (A), the Books Reviews in Arabic Dataset (B), and the Arabic Reviews dataset (C). A stacked-based ML model was built, which, when compared to other ML simple classifiers mentioned in this article, performed significantly better, and we used LR and SVM as meta-classifiers. Compared to existing machine learning simple classifiers, our stacked model with 10-fold cross-validation shows a higher accuracy, precision, Cohen's Kappa, recall, and F1-score.

In the following section, we will discuss some material that is pertinent to the backdrop. Section 3 describes our strategy. The results of the system's evaluation are covered in Section 4, which may be found here. In the final section, both a summary and some recommendations for the future are presented.

1.1 Problem statement and research questions

The main challenge of sentiment analysis is accurately categorizing sentiments expressed in a text, especially in the case of consumer sentiment analysis where opinions and attitudes are often nuanced and context-dependent. This research aims to address this challenge by developing a stacking machine learning model for consumer sentiment analysis using NLP techniques.

1.2 Contribution and significance of the study

The proposed model aims to enhance the efficiency and accuracy of sentiment analysis in consumer data by combining the advantages of different ML algorithms. The study contributes to the field of sentiment analysis by providing a comprehensive approach that incorporates NLP techniques and stacking machine learning models. The results of the study can benefit businesses, marketers, and customer experience managers in gaining valuable insights into their customers' attitudes and preferences.

2. RELATED WORK

In recent years, the number of algorithms and models capable of conducting sentiment analysis has increased with the growth of social networking and shopping online platforms. This is a direct consequence of the increase in the number of websites of this type. This part of the paper offers a thorough examination of the newest research carried out on SA. In more recent studies, approaches based on ML have been used to perform sentiment analysis in place of the more conventional methods.

A method for judging how people feel is employed to analyze datasets from Arabic social networks, as detailed in the study [14]. This approach was developed. The building of a corpus is carried out by hand using this method. The authors [15] present the results of the inquiry that was carried out. To determine whether or not the strategy given is accurate, a large number of ML techniques are utilized for the datasets used for training and the datasets used for testing. This helps to determine whether or not the approach is accurate. They [15] suggest a system that could be utilized for sentiment analysis by using sentiments produced from learning and teaching datasets. This system can better understand how people feel about different topics. This method can analyze how individuals think about many parts of the educational experience. The processed data are used as the basis for a feature selection, accomplished by developing four models. A support vector machine (SVM) approach is utilized in sentiment classification to supply the models with reliable findings. These two stages are necessary to obtain objective conclusions of the model [16], which can be accessed at this link and illustrate how an NB algorithm might be utilized to analyze Arabic tweets. In this technique of categorization, approaches dependent on the recurrence of phrases are applied. Following the division of the testing datasets into the five distinct components described above, the polarity of the tweets can then be characterized. Support vector machines, often known as SVMs and more frequently referred to by their acronym, are an additional method for determining an individual's attitude [17]. These machines are responsible for the categorization of the text datasets that were provided. After completing the five stages of the examination, the candidates' final scores are averaged to obtain an overall assessment of their performance. According to the data provided [18], Various ML algorithms are utilized to classify emotions, and the efficacy of the various ML strategies is evaluated.

The results of this research are described in the study [19], in which the authors address the development of an automatic classifier that uses techniques based on lexicons in conjunction with ML strategies. During this inquiry, the dataset is compiled and preprocessed, and a lexicon is built with the support of Senti Word-Net. The authors [20] determined the polarity of the feeling by analyzing the frequency with which various terms appeared within the dataset. This was done to classify the sensation that was felt. The dataset is divided into testing and training label columns so that different ML techniques can compare and contrast the effectiveness of their respective solutions. According to the findings presented [21], many strategies for feature extraction and sentiment analysis methods are given together in this passage. This post will investigate feature extraction techniques while utilizing various dataset domains to accomplish our goal. As can be seen [22], the SVM and NB classification approaches are put to use to classify the preferences of customers that were collected as part of the E-commerce dataset. The performance of ML techniques is studied, and the findings reveal that neural networks outperform support vector machines (SVM) in terms of their overall performance.

Mamun et al. [23] found that the ensemble approach (LR+RF+SVM) with frequency-inverse document frequency features outperformed other classifier models by 82%. This was determined by comparing the ensemble technique's accuracy to the other classifier models. This was determined by analyzing the results of a comparison between the accuracy of the ensemble technique and the accuracy of the other classifier models.

3. METHODOLOGY

In the following sentences, we will go into our ML models, which will include the datasets, feature extraction, and dataset processing. Nevertheless, before we get to that point, we will begin with the primary structure of our research, which is illustrated in Figure 1.

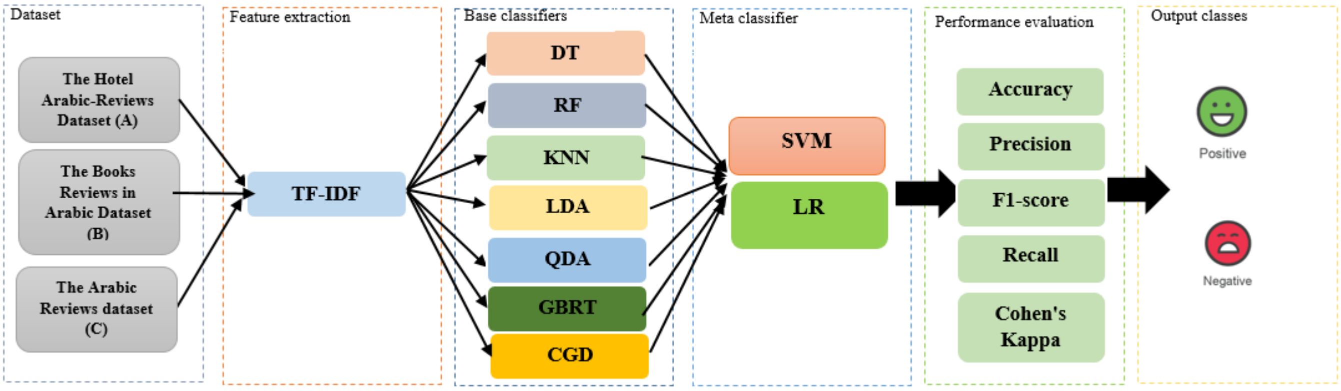

Figure 1. Global architecture of our model

3.1 Dataset used

For this paper, the following datasets were used to determine how well our model performed:

The Hotel Arabic-Reviews Dataset (A) [24] is made up of 490587 hotel reviews that were gathered from the Booking website throughout June and July of 2016. The evaluations are composed in dialectal Arabic (DA) and Modern Standard Arabic (MSA) and scored on a scale ranging from one to 10. The HARD dataset, which consisted of both positive and negative evaluations, was utilized in our analysis. The reviews were categorized as either positive (with scores between 4 and 5) or negative (with scores ranging from 1 to 3). The review types represented in this dataset include 93700 total reviews, with an equal number of negative (46850) and positive (46850) categories.

The Books Reviews in Arabic Dataset (B) [25] comprises 510,600 reviews of Arabic books obtained from GoodReads.com between June and July of 2016. The reviews are primarily written in DA and MSA style. We utilized the BRAD dataset, which includes equal good and negative testimonials and ratings. Only positive (four and five stars) and negative research were found in the reviews (one and two stars). The collection includes slightly more than 156,000 different reviews.

The Arabic Reviews dataset (C) [26] contains feedback from more than one hundred thousand customers on hotels, books, movies, and various other things, in addition to specific airlines. The reviewers' ratings are classified into negative, positive, or mixed. More than three reviewers gave it a positive rating, while fewer than three gave it an unfavorable rating. There is now a text and a label connected with each row, and the text in each row has been cleansed of non-Arabic characters and Arabic diacritics. There are only a few duplicate reviews in this collection.

3.2 Pre-processing

Pre-processing of datasets involves the processes performed to prepare raw data for machine learning tasks. These steps typically consist of data cleaning, data transformation, and data partitioning into testing and training sets. The specific pre-processing steps required vary depending on the form of data utilized and the objectives of the machine-learning task.

Some common pre-processing steps include:

• Tokenization, or segmentation, prepares a document for subsequent processing by separating it into a list of tokens consisting of numbers, words, and other special characters. This can be done by splitting documents, such as crawling reviews.

• The process of normalization converts all word tokens in a document to either all lowercase letters or all capital letters to normalize them. This is done because most assessments contain both uppercase letters and lowercase. With this method, one can improve the accuracy of their forecasts.

• The removal of typical stop words, such as prepositions, unnecessary words, memorable characters, and ASCII code, as well as excessive white spaces, new lines, and other elements, improves the functionality of the feature selection technique. This is accomplished through the use of the stop word removal technique.

• The stemming process involves converting all tokens into the root form, also known as the stem. This method is speedy and uncomplicated, and it makes the process of extracting characteristics much simpler.

3.3 Feature extraction

To accomplish our objectives in natural language processing, we make use of a variety of machine-learning strategies. During the training process for our model, we extract each phrase's features based on two primary criteria. To put this strategy into action, we use a method known as tokenization, which involves deconstructing phrases into the individual words that make them up, irrespective of how frequent or uncommon the individual terms may be. In addition, we use a numerical figure that determines the significance of a phrase in a text. This technique is TF-IDF. Because of the method's track record of success for various languages, we have decided to embrace it. As a direct consequence, our machine learning algorithms have accomplished remarkable things.

3.4 Machine Learning (ML) models: An overview

ML uses various sequences to learn from data and assign it to categories. A categorization model is formed utilizing the training set's knowledge to categorize input data as either positive or negative [27]. The training set contains the names and categories of the input feature vectors. The extracted feature sets are used in training the classifier, which determines whether the analysis of the dataset is favorable or unfavorable based on the results. The objective of the stacking techniques in machine learning is to produce a more accurate prediction model by integrating several base models.

3.4.1 Support Vector Machine (SVM)

SVM is a popular supervised ML algorithm for regression and classification analysis. In the Support vector machine, the objective is to identify the hyperplane that most effectively separates the data into distinct classes. The hyperplane is chosen to maximise the margin between the different classes. The points closest to the hyperplane are known as support vectors, which play a critical role in determining the hyperplane. SVM can be used for both non-linearly and linearly separable data and can handle high-dimensional data. A powerful algorithm can provide accurate results with relatively small datasets. SVM can also handle imbalanced datasets, where the number of examples in one class is much larger than the other [28].

3.4.2 Logistic regression (LR)

LR is a prominent statistical learning approach used to predict a binary output variable from one or more input variables or predictors. The outcome variable is typically coded as 0 or 1, indicating the presence or absence of a certain event or condition [29].

In LR, the result is modeled using a logistic function or sigmoid function, which maps any real-valued input to the range [0,1]. The logistic function calculates the probability of the positive class given the input variables, and the negative class probability is simply the complement of the positive class probability. The logistic regression algorithm estimates the model parameters by maximizing the likelihood function of the data, which is the probability of observing the data given the model parameters.

3.4.3 Decision Tree (DT)

DT is a model that may be applied to classification and regression problems. The data is divided into subsets using a recursive process that is based on the value of one of the features. This process continues until the subsets only include one class or until a set of specified stopping criteria is satisfied. Overfitting is a risk associated with decision trees; however, this issue can be mitigated by employing strategies such as pruning and establishing a limited depth [30].

3.4.4 Random Forest (RF)

The RF is an ensemble learning technique comprising several different decision trees. The algorithm chooses a subset of features for each tree based on a random selection, and then it aggregates the predictions from all of the trees to develop the overall forecast. Random forests are frequently utilized when compared to a single decision tree because they are better at preventing overfitting and improving accuracy [31].

3.4.5 K-Nearest Neighbors (KNN)

KNN is a non-parametric technique applicable to regression and classification applications. Finding the k data points that are geographically closest to a new data point is the first step in the algorithm's process. Then, the classes of those data points are utilized to guess the nature of the unique data point. The value of k can be determined based on the cross-validation results; however, choosing the best number can be time-consuming [32].

3.4.6 Linear Discriminant Analysis (LDA)

LDA is an example of a supervised ML approach that may be used for classification. This type of algorithm searches for a linear combination of features that can best separate the classes. It assumes that the data follow a normal distribution and that the covariances of the different types are equivalent. Because it projects the data onto a lower-dimensional space while preserving the information that allows for maximum discrimination, LDA is frequently utilized in the dimensionality reduction process in feature extraction [33].

3.4.7 Quadratic Discriminant Analysis (QDA)

QDA stands for quadratic discriminant analysis. It is comparable to LDA, except it enables individualized covariances for each class. This results in QDA being a more flexible model but requiring additional data to estimate the covariance matrices accurately. When the underlying distribution of the data is nonlinear, QDA can provide more accurate results than LDA [34].

3.4.8 Gradient Boosting Regression Trees (GBRT)

GBRT [35] is an approach for supervised learning used for regression and classification. It is an ensemble method that makes a final forecast by combining the outcomes of multiple DTs. The GBRT sequentially creates the trees, with each new tree attempting to repair the mistakes that were introduced by the trees that came before it. The ultimate forecast is derived from a weighted sum of the predictions made by each tree. The GBRT algorithm is reliable and capable of dealing with nonlinear connections between the characteristics and the target variable. Nevertheless, if the parameters are incorrectly calibrated, it may overfit the data.

3.4.9 Stochastic Gradient Descent (SGD) Classifier

SGD Classifier stands for stochastic gradient descent, and it is an example of a supervised machine learning approach that is used for categorization. An optimization method will make changes to the model parameters in an iterative manner. These changes will be based on a batch, a randomly selected chunk of the training data. SGD can be sensitive to the initial learning rate and the batch size, even though it is computationally efficient and can manage big datasets. The SGD classifier can solve linear and nonlinear problems and is compatible with a wide variety of loss functions, including logistic regression and support vector machines [36].

3.4.10 Ensemble stacking ML model

Stacking involves using heterogeneous weak classifiers trained in parallel and integrated by a meta-learner to give a forecast based on the projections of the individual susceptible learners. As can be seen in Figure 2, the meta-learner utilizes the predictions as input features and the goal as ground truth values. It then learns how to optimally combine these input predictions to generate a more accurate output prediction. This method is efficient because the various models can pick up knowledge that complements what they already know, ultimately enhancing performance. Stacking is a beneficial technique when working with datasets that are not evenly distributed because it lowers the prediction variance. It is crucial, however, to carefully calibrate both the separate models and how they are integrated. This is because, when compared to other machine learning techniques, it is a sophisticated approach.

This research used DT, RF, KNN, LDA, QDA, GBRT, and SGD as base classifiers. SVM and LR as meta-classifiers.

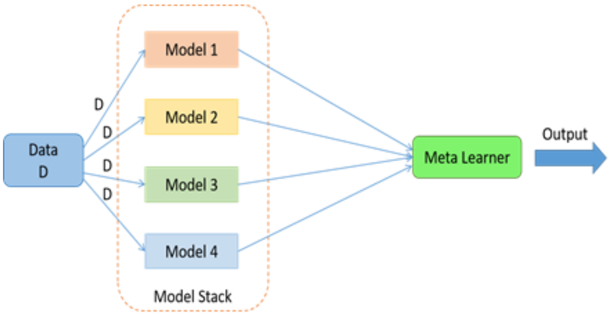

Figure 2. The stacking model

3.5 Performance measures

In this section, we will discuss the assessment measures we employed to evaluate the performance of our model.

For this discussion, the terms "false positives," "true positives," "false negatives," and "true negatives," respectively, are abbreviated as Fp, Tp, Fn, and Tn, as shown in Table 1 [2, 31-33].

Table 1. Summary of sentiment analysis evaluation metrics

Performance Measure	Description	Calculation
Accuracy	The proportion of correctly classified instances among all instances in the dataset.	$Acc = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$
Precision	The proportion of true positives (correctly classified instances of the positive class) among all instances predicted as positive.	$Pre = \frac{Tp}{Tp + Fp}$
F1-score	The harmonic mean of precision and recall provides a balanced measure of the classifier's performance.	$\frac{2 * (\text{Precision. Recall})}{(\text{Precision} + \text{Recall})}$
Recall (Sensitivity)	The proportion of true positives among all instances of the positive class.	$\frac{Tp}{Tp + Fn}$
Cohen's Kappa [37]	A statistic that measures the agreement between predicted and true classifications, taking into account the expected agreement due to chance. Where Pa represents the raters' actual agreement and Pb their probability of agreement.	$\text{Kappa} = \frac{\text{Pa} - \text{Pb}}{1 - \text{Pb}}$

4. EXPERIMENTS AND FINDINGS

Using the datasets mentioned above, we evaluated the performance of our recommended machine learning model based on stacking many classifiers. To determine how well it performed, we compared its results to those of several different individual classifiers, including DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier. We used metrics such as recall, precision, Cohen's Kappa, F1-score, and accuracy in evaluating the model's performance. The results of our investigation are summarized in the tables that may be found below.

On dataset A, as shown in Table 2, we utilized 5-fold cross-validation (cv) on dataset A to evaluate the performance of various classifiers, including DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier. The resulting accuracies were 79.53%, 80.43%, 78.65%, 80.02%, 84.92%, 79.53%, and 84.43%, respectively. Our stacked model achieved the highest accuracy of 88.25%, surpassing the accuracy of all individual classifiers.

We performed 10-fold cross-validation on dataset A and obtained the accuracy results for each classifier, including DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier. The accuracies obtained were 80.25%, 81.16%, 79.37%, 80.75%, 85.69%, 80.25%, and 85.20%, respectively. However, our stacked machine learning model achieved the highest accuracy of 91.46% according to Table 3.

Table 2. The model's evaluation on Dataset A with k-fold validation (k=5)

Classifier	TF-IDF (k=5)				
	Precision	Accuracy	Cohen's Kappa	F1-Score	Recall
DT	79.37	79.53	75,81	76,35	73,39
RF	79.26	80.43	71,48	81,05	77,91
KNN	78.49	78.65	74,97	75,50	72,58
LDA	79.86	80.02	73,28	76,82	73,84
QDA	84.75	84.92	72,95	81,52	78,36
GBRT	79.37	79,53	75,81	76,35	73,39
SGD	80.26	84,43	74,48	78,05	77,91
Stacked model	88.07	88.25	81.12	83.71	81.44

Table 3. The model's evaluation on dataset A with k-fold validation (k=10)

Classifier	TF-IDF (k=10)				
	Precision	Accuracy	Cohen's Kappa	F1-Score	Recall
DT	80.09	80.25	76.50	77.04	74.06
RF	79.98	81.16	72.13	81.79	78.62
KNN	79.20	79.37	75.65	76.19	73.24
LDA	80.59	80.75	73.95	77.52	74.51
QDA	85.52	85.69	73.61	82.26	79.07
GBRT	80.09	80.25	76.50	77.04	74.06
SGD	80.99	85.20	75.16	78.76	78.62
Stacked model with LR	88.21	89.73	79.41	83.96	82.71
Stacked model with SVM	90.27	91.46	80.15	85.80	83.48

Table 4. The model's evaluation on dataset B with k-fold validation (k=5)

Classifier	TF-IDF (k=5)				
	Precision	Accuracy	Cohen's Kappa	F1-Score	Recall
DT	77.74	77.89	74.26	74.78	71.88
RF	82.53	82.69	78.83	79.38	76.31
KNN	76.88	77.03	73.43	73.95	71.09
LDA	78.22	78.37	74.71	75.24	72.33
QDA	83.01	83.17	79.29	79.84	76.75
GBRT	77.74	77.89	74.26	74.78	71.88
SGD	82.53	82.69	76.87	76.44	76.31
Stacked model with LR	85.31	84.71	78.01	81.81	77.51
Stacked model with SVM	86.26	86.43	79.46	82.97	79.76

We performed 5-fold cross-validation on dataset B for each classifier, including DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier. The resulting accuracies were 77.89%, 82.69%, 77.03%, 78.37%, 83.17%, 77.89%, and 82.69%, respectively. However, our stacked machine learning model achieved the highest accuracy of 86.43% as per Table 4.

Table 5. The model's evaluation on Dataset B with k-fold validation (k=10)

Classifier	TF-IDF (k=10)				
	Precision	Accuracy	Cohen's Kappa	F1-Score	Recall
DT	78.31	78.46	74.80	75.32	72.41
RF	83.13	83.29	79.40	79.96	76.87
KNN	77.44	77.59	73.97	74.49	71.61
LDA	78.79	78.95	75.26	75.79	72.85
QDA	83.61	83.78	79.86	80.42	77.31
GBRT	78.31	78.46	74.80	75.32	72.41
SGD	83.13	83.29	77.43	77.00	76.87
Stacked model with LR	85.41	85.25	78.86	81.47	78.91
Stacked model with SVM	86.89	87.06	80.04	83.58	80.34

We performed 10-fold cross-validation on dataset B and obtained the accuracy results for each classifier, including DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier. The accuracies obtained were 78.46%, 83.29%, 77.59%, 78.95%, 83.78%, 78.46%, and 83.29%, respectively. However, our stacked machine-learning model achieved the highest accuracy of 87.06% as shown in Table 5.

Table 6. The model's evaluation on Dataset C with k-fold validation (k=5)

Classifier	TF-IDF (k=5)				
	Precision	Accuracy	Cohen's Kappa	F1-Score	Recall
DT	77.10	77.25	73.65	74.16	71.29
RF	81.85	82.01	78.18	78.73	75.68
KNN	76.25	76.40	72.83	73.34	70.50
LDA	77.58	77.72	74.10	74.62	71.73
QDA	82.33	82.49	78.64	79.18	76.12
GBRT	77.10	77.25	73.65	74.16	71.29
SGD	81.85	82.01	76.24	75.81	75.68
Stacked model with LR	84.32	84.21	75.96	81.01	77.98
Stacked model with SVM	85.55	85.72	78.81	82.29	79.10

On dataset C, according to Table 6, we conducted 5-fold cross-validation (cv) for each classifier; DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier, we obtained accuracy of 77.25%, 82.01%, 76.40%, 77.72%, 82.49%, 77.25%, 82.01% respectively and achieved more results in terms of accuracy for our stacked machine learning model as 85.72%.

Table 7. The model's evaluation on dataset C with k-fold validation (k=10)

Classifier	TF-IDF (k=10)				
	Precision	Accuracy	Cohen's Kappa	F1-Score	Recall
DT	77.96	78.11	74.46	74.98	72.09
RF	82.76	82.92	79.04	79.60	76.53
KNN	77.09	77.24	73.64	74.16	71.29
LDA	78.44	78.60	74.92	75.45	72.52
QDA	83.24	83.40	79.50	80.06	76.96
GBRT	77.96	78.11	74.46	74.98	72.09
SGD	82.76	82.92	77.08	76.66	76.53
Stacked model with LR	85.35	85.13	78.21	82.57	78.11
Stacked model with SVM	86.50	86.67	79.68	83.21	79.98

For 10-fold cross-validation, we obtained for every classifier; DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier, we obtained accuracy as 78.11%, 82.92%, 77.24%, 78.60%, 83.40%, 78.11%, 82.92% respectively and achieved more results in terms of accuracy for stacked machine learning model as 86.67% as per Table 7.

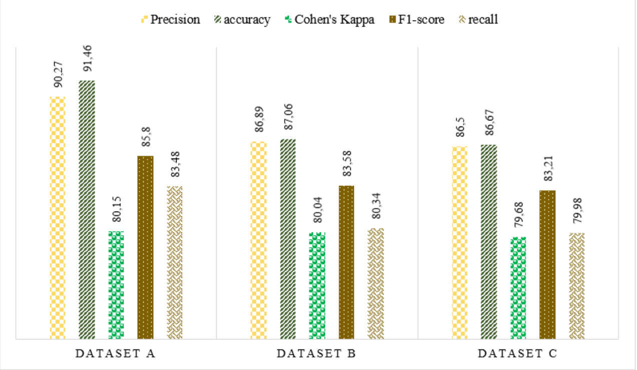

Figure 3. Our model's performance with k=5

Because we have access to the necessary experimental data, we can investigate the degree to which the performance of our stacked machine-learning model is correlated with that of the individual classifiers. The acquired results show how well each classifier performed regarding the five metrics used for evaluation. In this part, a comprehensive review of the effectiveness of the suggested approach is provided. According to the assessment criteria for Dataset A, B, and C with k=5 and k=10, as shown in Figures 3 and 4, our stacked machine learning model achieves commendable results. These results are depicted in these figures.

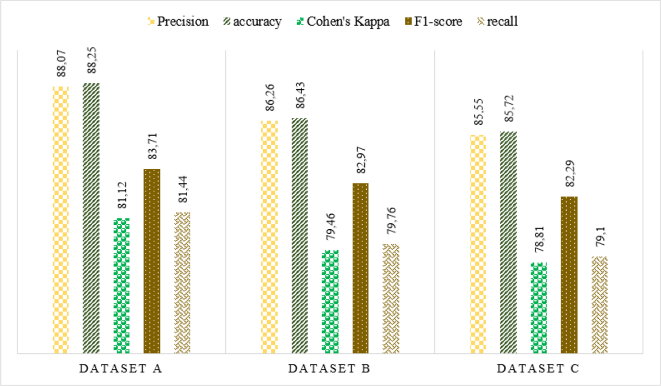

Figure 4. Our model's performance with k=10

4.1 Discussion

We tested our proposed machine learning model using several classifiers in this investigation. We compared its recall, precision, Cohen's Kappa, F1-score, and accuracy to DT, RF, KNN, LDA, QDA, GBRT, and SGD classifiers.

We tested the classifiers on dataset A using 5-fold cross-validation. DT, RF, KNN, LDA, QDA, GBRT, and SGD classifiers have an accuracy of 79.53%, 80.43%, 78.65%, 80.02%, 84.92%, and 84.43%. Our stacked model outperformed all classifiers with 88.25% accuracy. The stacked model's ensemble of classifiers outperformed each classifier on dataset A in accuracy.

We next did a 10-fold cross-validation on dataset A to determine classifier accuracy. DT, RF, KNN, LDA, QDA, GBRT, and SGD classifiers have 80.25%, 81.16%, 79.37%, 80.75%, 85.69%, and 85.20% accuracy. Again, our layered machine learning model had the greatest accuracy of 91.46%. These results reinforce the stacked model's accuracy advantage on dataset A.

Each classifier underwent 5-fold cross-validation on dataset B. DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier have accuracies of 77.89%, 82.69%, 77.03%, 78.37%, 83.17%, and 82.69%. However, our stacked model was most accurate at 86.43%. In the 10-fold cross-validation on dataset B, DT, RF, KNN, LDA, QDA, GBRT, and SGD Classifier had accuracies of 78.46%, 83.29%, 77.59%, 78.95%, 83.78%, and 83.29%. Again, our layered machine-learning model had a maximum accuracy of 87.06%. Both datasets (A and B) show that the stacked model performs better.

We also tested the classifiers on dataset C using 5-fold cross-validation. DT, RF, KNN, LDA, QDA, GBRT, and SGD classifiers have accuracy of 77.25%, 82.01%, 76.40%, 77.72%, 82.49%, and 82.01%. Our layered machine learning model improved classification accuracy on dataset C with 85.72% accuracy.

All datasets showed that the stacked machine learning model outperformed the separate classifiers in accuracy. This suggests that combining classifier predictions improves classification performance. The stacked model's greater accuracy suggests practical applications in numerous categorization fields.

Our study shows that the proposed stacked machine learning approach improves classification accuracy over individual classifiers. Multiple datasets show that the layered model performs better, suggesting practical applications. These findings can be used to test the stacking model in additional domains and datasets.

5. CONCLUSION

Arabic Sentiment Analysis (ASA) has emerged as an essential component in various fields, including production, politics, and service provision. Arabic material featuring people's perspectives on different themes of considerable interest to ASA academics may be abundant on social networks (SN). The processing of data in the ASA area is significantly aided by applying machine-learning techniques. In this regard, seven different classifier models, such as Random Forest (RF), Decision Tree (DT), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Gradient Boosting Regression Trees (GBRT), and Stochastic Gradient Descent (SGD) Classifier, were utilized to categorize a collection of comments and reviews as being either positive or negative. Our stacked model-based machine learning algorithms were evaluated alongside these models for comparison. The research started by performing some preliminary work on three database files. The effectiveness of the used classifiers in this research was evaluated using five different measures, including precision, kappa, accuracy, F1-score, and recall. A 10-fold cross-validation ensemble classifier performed superiorly to all other models while testing ensemble-based sentiment classification. This was the case across all evaluation criteria. The findings suggest that ensemble-based sentiment classification can significantly enhance the accuracy and reliability of ASA, which can be leveraged in various applications, including market analysis, political sentiment tracking, and customer feedback analysis.

In further research, we intend to push the AraBERT model's ability to improve ASA by applying it to an extensive database containing comments and reviews in various Arabic dialects and standard Arabic languages.

REFERENCES

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

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

[3] Almuraqab, N.A.S. (2021). Determinants that influence consumers' intention to purchase smart watches in the UAE: A case of university students. Advances in Science, Technology and Engineering Systems Journal, 6(1): 1249-1256. https://doi.org/10.25046/aj0601142

[4] Al Shamsi, A.A., Abdallah, S. (2021). Text mining techniques for sentiment analysis of Arabic dialects: Literature review. Advances in Science, Technology and Engineering Systems Journal, 6(1): 1012-1023. https://doi.org/10.25046/aj0601112

[5] Hicham, N., Karim, S. (2022). Machine learning applications for consumer behavior prediction. In The Proceedings of the International Conference on Smart City Applications, pp. 666-675. https://doi.org/10.1007/978-3-031-26852-6_62

[6] Gregoriades, A., Pampaka, M., Herodotou, H., Christodoulou, E. (2021). Supporting digital content marketing and messaging through topic modeling and decision trees. Expert Systems with Applications, 184: 115546. https://doi.org/10.1016/j.eswa.2021.115546

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

[8] Sánchez-Rada, J.F., Iglesias, C.A. (2019). Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. Information Fusion, 52: 344-356. https://doi.org/10.1016/j.inffus.2019.05.003

[9] Choi, Y., Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. Information Systems Frontiers, 19: 993-1012. https://doi.org/10.1007/s10796-017-9741-7

[10] Valdivia, A., Luzón, M.V., Cambria, E., Herrera, F. (2018). Consensus vote models for detecting and filtering neutrality in sentiment analysis. Information Fusion, 44: 126-135. https://doi.org/10.1016/j.inffus.2018.03.007

[11] Birjali, M., Beni-Hssane, A., Erritali, M. (2016). Measuring documents similarity in large corpus using MapReduce algorithm. In 2016 5th International Conference on Multimedia Computing and Systems (ICMCS), Marrakech, Morocco, pp. 24-28. https://doi.org/10.1109/ICMCS.2016.7905587

[12] Ramírez-Tinoco, F.J., Alor-Hernández, G., Sánchez-Cervantes, J.L., Olivares-Zepahua, B.A., Rodríguez-Mazahua, L. (2018). A brief review on the use of sentiment analysis approaches in social networks. In Trends and Applications in Software Engineering, pp. 263-273. https://doi.org/10.1007/978-3-319-69341-5_24

[13] Mäntylä, M.V., Graziotin, D., Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27: 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

[14] Hnaif, A.A., Kanan, E., Kanan, T. (2021). Sentiment analysis for Arabic social media news polarity. Intelligent Automation & Soft Computing, 28(1): 107-119. https://doi.org/10.32604/iasc.2021.015939

[15] Muhammad, A., Abdullah, S., Sani, N.S. (2021). Optimization of sentiment analysis using teaching-learning based algorithm. Computers, Materials & Continua, 69(2): 1783-1799. https://doi.org/10.32604/cmc.2021.018593

[16] Alwakid, G., Osman, T., Haj, M.E., Alanazi, S., Humayun, M., Sama, N.U. (2022). MULDASA: Multifactor lexical sentiment analysis of social-media content in nonstandard Arabic social media. Applied Sciences, 12(8): 3806. https://doi.org/10.3390/app12083806

[17] Habbat, N., Anoun, H., Hassouni, L. (2022). Sentiment analysis and topic modeling on Arabic Twitter data during COVID-19 pandemic. Indonesian Journal of Innovation and Applied Sciences (IJIAS), 2(1): 60-67. https://doi.org/10.47540/ijias.v2i1.432

[18] Wongkar, M., Angdresey, A. (2019). Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter. In 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, pp. 1-5. https://doi.org/10.1109/ICIC47613.2019.8985884

[19] Jose, R., Chooralil, V.S. (2016). 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), Ernakulam, India, pp. 64-67. https://doi.org/10.1109/SAPIENCE.2016.7684133

[20] Poornima, A., Priya, K.S. (2020). A comparative sentiment analysis of sentence embedding using machine learning techniques. In 2020 6th international conference on advanced computing and communication systems (ICACCS), Coimbatore, India, pp. 493-496. https://doi.org/10.1109/ICACCS48705.2020.9074312

[21] Habib, M.W., Sultani, Z.N. (2021). Twitter sentiment analysis using different machine learning and feature extraction techniques. Al-Nahrain Journal of Science, 24(3): 50-54. https://doi.org/10.22401/ANJS.24.3.08

[22] Mamatha, M., Thriveni, J., Venugopal, K.R. (2018). Techniques of sentiment classification, emotion detection, feature extraction and sentiment analysis a comprehensive review, International Journal of Computer Sciences and Engineering, 6(1): 244-261. https://doi.org/10.26438/ijcse/v6i1.244261

[23] Mamun, M.M.R., Sharif, O., Hoque, M.M. (2022). Classification of textual sentiment using ensemble technique. SN Computer Science, 3(1): 49. https://doi.org/10.1007/s42979-021-00922-z

[24] Elnagar, A., Khalifa, Y.S., Einea, A. (2018). Hotel Arabic-reviews dataset construction for sentiment analysis applications. Intelligent natural language processing: Trends and Applications, 740: 35-52. https://doi.org/10.1007/978-3-319-67056-0_3

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

[26] Vincent, R., Bhatia, P., Rajesh, M., Sivaraman, A.K., Al Bahri, M.S.S. (2020). Indian currency recognition and verification using transfer learning. International Journal of Mathematics and Computer Science, 15(4): 1279-1284. https://doi.org/10.6084/M9.FIGSHARE.16944082

[27] Hashim, N.N.W.N., Basri, N.A., Ezzi, M.A.E.A., Hashim, N.M.H.N. (2022). Comparison of classifiers using robust features for depression detection on Bahasa Malaysia speech. IAES International Journal of Artificial Intelligence (IJ-AI), 11(1): 238-253.

https://doi.org/10.11591/ijai.v11.i1.pp238-253

[28] Hicham, N., Karim, S., Habbat, N. (2023). Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach. International Journal of Electrical and Computer Engineering (IJECE), 13(4): 4504-4515. https://doi.org/10.11591/ijece.v13i4.pp4504-4515

[29] Tangwannawit, S., Tangwannawit, P. (2022). An optimization clustering and classification based on artificial intelligence approach for internet of things in agriculture. IAES International Journal of Artificial Intelligence, 11(1): 201-209. https://doi.org/10.11591/ijai.v11.i1.pp201-209

[30] Hicham, N., Karim, S. (2022). Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering. International Journal of Advanced Computer Science and Applications, 13(10): 122-130, https://doi.org/10.14569/ijacsa.2022.0131016

[31] Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems, 29(5): 1774-1785. https://doi.org/10.1109/TNNLS.2017.2673241

[32] Tharwat, A., Gaber, T., Ibrahim, A., Hassanien, A.E. (2017). Linear discriminant analysis: A detailed tutorial. AI Communications, 30(2): 169-190. https://doi.org/10.3233/AIC-170729

[33] Ghojogh, B., Crowley, M. (2019). Linear and quadratic discriminant analysis: Tutorial. arXiv preprint arXiv:1906.02590. https://doi.org/10.48550/arXiv.1906.02590

[34] Prettenhofer, P., Louppe, G. (2014). Gradient boosted regression trees in scikit-learn. In PyData 2014, London, United Kingdom.

[35] Newton, D., Yousefian, F., Pasupathy, R. (2018). Stochastic gradient descent: Recent trends. Recent Advances in Optimization and Modeling of Contemporary Problems, 193-220. https://doi.org/10.1287/educ.2018.0191

[36] Warrens, M.J. (2015). Five ways to look at Cohen's kappa. Journal of Psychology & Psychotherapy, 5(4): 1000197. https://doi.org/10.4172/2161-0487.1000197.

[37] Vergni, L., Todisco, F., Di Lena, B. (2021). Evaluation of the similarity between drought indices by correlation analysis and Cohen's Kappa test in a Mediterranean area. Natural Hazards, 108(2): 2187-2209. https://doi.org/10.1007/s11069-021-04775-w