An Improved Arabic Sentiment Analysis Approach using Optimized Multinomial Naïve Bayes Classifier

Ahmed Alsanad
STC's Artificial Intelligence Chair, Department of Information Systems
College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

***Abstract*—Arabic sentiment analysis has emerged during the last decade as a computational process on Arabic texts for extracting people's attitudes toward targeted objects or their feelings and emotions regarding targeted events. Sentiment analysis (SA) using machine learning (ML) methods has become an important research task for developing various text-based applications. Among different ML classifiers, multinomial Naïve Bayes (MNNB) classifier is widely used for documents classification due to its ability for performing statistical analysis of text contents. It significantly simplifies textual-data classification and offers an alternative to heavy ML-based semantic analysis methods. However, the MNNB classifier has a number of hyper-parameters affects the classification task of texts and controls the decision boundary of the model itself. In this paper, an optimized MNNB classifier-based approach is proposed for improving Arabic sentiment analysis. A number of experiments are conducted on large sets of Arabic tweets to evaluate the proposed approach. The optimized MNNB classifier is trained on three datasets and tested on a different separated test set to show the performance of developed approach. The experimental results on the test set revealed that the optimized MNNB classifier of proposed approach outperforms the traditional MNNB classifier and other baseline classifiers. The accuracy rate of the optimization approach is increased by 1.6% compared with using the default values of the classifier's hyper-parameters.**

Keywords*—Machine learning; Arabic sentiment analysis; optimized multinomial Naïve Bayes (MNNB) classifier; hyper-parameters optimization

I. Introduction

Recently, sentiment analysis (SA) using machine learning methods has become an important research task for developing various text-based applications [1, 2]. It is used in natural language processing (NLP), text search, and computational linguistics for extracting people's opinions or feelings about products, events, or other in one way or another [3]. The SA imposes the identification of several elements which are four elements including the entity, its aspect, the opinion holder, and its feeling and can categorize the opinions extracted into either a subjective or objective text. The subjective text can also be categorized into negative or positive feelings. Several methods have been conducted on sentiment analysis of several languages, including Arabic, and differences have been discovered, as NLP in Arabic is still in its early steps and lacks tools and resources [4]. Hence, the Arabic language still faces difficulties in NLP tasks because of its complexity, structure, and the different dialects to which it belongs. A large number of tools and methods have been used, in the literature, to perform the sentiment analysis task. Most of them are considered in English, which is the science language, these tools and methods are based on either a machine learning approach or a semantic approach [5]. The semantic approach extracts emotion words and computes their poles based on the emotion dictionary. Conversely, in a machine learning approach to building a new model, machine learning classifiers are trained on pre-labeled data after transforming into feature vectors. Finally, the new model can be applied for predicting a new category of data based on these characteristics. It should be noted that these methods can be modified to another language, such as Arabic. The Arabic language did not receive the effort that other languages did [6]. However, several studies of sentiment analysis of Arabic writing have been proposed [7]. A decade ago, Arabic sentiment analysis became one of the most common information mining forms in many fields. These analytics have contributed to achieve many benefits, such as offering brand value insights for a product or service Invite potential product customers, identify social media influencers, and detect spam. Consequently, Arabic sentiment analysis has been investigated in different contexts, and a number of techniques in several studies have been published on this topic. However, there is still a limitation in MNNB classifier, which is widely used to analyze users' sentiment from texts and classify their topics.

This study aims to develop an optimized simple ML model-based approach for improving Arabic sentiment analysis using grid search algorithm and multinomial naïve Bayes (MNNB) classifier. The approach is able to select the optimal values for alpha hyper-parameter and control the decision boundary of the classifier. Through this proposed approach, the research contributions to the field can be presented in the following points:

- Propose an improved approach for Arabic sentiment analysis using a grid search algorithm and MNNB classifier that can be able to optimize the learning process of the classifier.
- Use the grid search algorithm as a selection step to assign the optimal or near-optimal values of MNNB classifier's alpha-parameter, improving the sentiment analysis of Arabic tweets.
- Train the proposed approach on three datasets with a large number of Arabic tweets for achieving the diversity in learning the ML models.

- Evaluate and compare the performance results of the optimized model with the baseline models on the same test dataset.

The rest of the paper is structured as follows: Section II gives a detailed literature review for the study. Section III describes the research methodology. Section IV presents the experimental results with discussions and findings. Finally, the conclusions and future work are summarized in Section V.

II. Literature Review

Nowadays, the rapid development of social media makes the text messages posted by users become the largest public data source in the world. Such text messages contain an important information and a great commercial and research value. Sentiment analysis and text analytics using NLP is one of the key methods that can provide a necessary support for text analysis in the social media. Consequently, text analytics technique including sentiment analysis, entity recognition, topic modeling, and text summarization using NLP in social media has attracted widespread attention.

The social media Arabic sentiment analysis can analyze and mine the tendency and view of user expressions from his subjective text. This analyzing supports the decision making of different researchers, users, government agencies, and business organizations. In this context, the worth question of discussion is how to effectively mine these massive textual information, identify the sentiment in it, and use it reasonably and effectively. Sentiment analysis, also known as propensity analysis, is a computational study of emotions, opinions, and feelings held by people about things and their attributes [8]. Things can be services, products, individuals, organizations, questions, events, or topics. Sentiment analysis task can also be defined as the process of automatically mining attitudes, opinions, opinions, and emotions from speech, text, Weibo, and other data sources through NLP technology [9]. Text sentiment analysis is to analyze the sentiment of a paragraph of text, as the basic work of public opinion monitoring, and has a wide range of uses.

Social networks are getting more and more popular, and "opinion leaders" are getting more and more. Sites that allow users to rate product and service evaluations have sprung up, and user reviews and suggestions can be spread throughout the network. These text-type data are undoubtedly the source of the power of precision marketing. Enterprises can build their own digital image based on sentiment analysis, identify new market opportunities, do a good job of market segmentation, and then promote the successful listing of products. But grasping the value of these reviews is also a huge challenge for companies. Governments, like enterprises, need to monitor, alleviate, and lead public opinion through sentiment analysis, and eliminate social conflicts. The above is the application background of sentiment analysis. But contrary to such an important background is the weakness of the Arabic sentiment analysis system.

Common sentiment analysis is separated into dictionary-based sentiment analysis and supervised model-based analysis. Dictionary-based sentiment-analysis, as the name suggests, relies heavily on the construction of sentiment dictionaries. Ku et al. [10] and Kaji et al. [11] conducted in-depth research on the construction of sentiment dictionaries. Generally, the emotional words are first divided into positive (meaning) and negative (derogatory), and then the number of positive words and negative words of an Arabic text to be analyzed is counted. If the positive words number is greater than the negative number of words, this text belongs to the positive emotion, otherwise it belongs to the negative emotion.

Some researchers have artificially weighted sentiment dictionaries. However, no matter how it is changed, this analysis method has the following limitations: first, the accuracy is very low, which can hardly support the requirements of public opinion monitoring; second, the positive or negative tendency or weight of emotional words is manually defined, and the workload is huge and very arbitrary; in the end, this approach is almost ineffective for negative sentences and sentences reinforced by adverbs of degree, thus losing the ability to analyze the delicateness (degree) of emotions.

Social media is an online interactive platform based on online social networks and the main form of Internet users' creation and dissemination of information. Twitter and Facebook are typical examples of online social networks. The emergence of social media has epoch-making significance, so that the general public's emotions can be easily and fully expressed, spread, and influence each other. There are a large number of research problems in the field of sentiment analysis. Many of today's naming related tasks with small differences are usually included in the research field of sentiment analysis, such as opinion mining, opinion, sentiment analysis, comment mining and so on. Text sentiment analysis [12-15] aims to analyze the attitudes and sentiment of opinion expressers, that is, to analyze the subjective information in the text.

Although text sentiment analysis studies texts with a polarity have started before the year 2000, few scholars studied the sentiment from texts in the field of linguistics and NLP. This may be partly because there was no growth in digital records at that time. With the explosive growth of the Internet and social media, people can have an uninterrupted data stream and store it in digital form, which is also an important reason why sentiment analysis has maintained a consistent growth rate with the Internet in recent years. For many years, social media systems have provided users with a very convenient channel to communicate and share. The important carrier of information is text information in social media.

People are keen to be in such a free and convenient environment that is not limited by time and space. Make your own voice, express your views on everything, and establish connections between users. This user-generated content provides researchers with great convenience to track, collect, store, retrieve, and analyze people's emotional changes. Appearance has injected new vitality into sentiment analysis, and sentiment analysis has also provided a new research area for social media analysis.

As early as 1997, Tiwari et al. [16] began to try to use the conjunctions in linguistics that are binding on sentiment words to infer the opinions and attitudes of the whole article, that is, to use adjectives with known sentiments to infer The emotional

tendency of an adjective. Turney et al. [17] used the association between words and some seed words with obvious semantic tendencies and combined statistical methods to identify the propensity characteristics of words. The research term of sentiment analysis may be first proposed in the literature of Nasukawa and Yi [14] in 2003, but many related work of sentiment analysis has been started before 2003 [15-17]. Earlier related studies include interpretation of metaphors, extraction of adjectives with emotions, sentiment calculation, subjective analysis, and so on.

The existing applications and research of sentiment analysis are mainly focused on text, which has become a hot spot in the field of NLP research. Since 2002, sentiment analysis research has become very active. In addition to the large number of trending texts in social media, its extensive application scenarios have become increasingly prominent in various human activities. This is also social media sentiment analysis is different from traditional text sentiment analysis. Social sentiment-oriented text sentiment analysis is mainly based on information sharing and interactive review mechanisms. When people need to make a decision at a certain time, most people often refer to the opinions of others. This situation is not only for individuals but also for enterprises and institutions.

Because the large amount of information with user sentiment is publicly available on the Internet, companies no longer use a large number of questionnaires to collect and understand consumer opinions on their products, and the government can easily grasp the public's perspective to supervise their regions. Therefore, social media from about 2006, sentiment analysis has ushered in its prosperity. Its widespread application has given rise to a strong demand for research, and at the same time brought many unprecedented challenges. These challenges are exactly the problems that this article needs to solve.

The related research on text sentiment analysis for social media has been widely concerned by academia and industry. Researchers and institutions have invested a lot of manpower and material resources in order to use text sentiment analysis technology to obtain relevant information. At present, text sentiment analysis has been discussed in many international top conferences.

In the industrial sector, such as major e-commerce shopping websites and portals, they have applied sentiment analysis technology to user review analysis, and found problems in the product and improved them through user reviews, thereby achieving the goal of increasing product sales.

Develop a comprehensive social media analytics tool to help decision makers with external customers to understand sentiment and feedback mapped to the services/products in discussion. The data source is social media feeds such as Twitter or Facebook. The model should be an engine to analyze Saudi dialect and English using advanced NLP algorithms which can be used with any free text platforms to understand main topics and sentiment analysis.

In recent years, Arabic sentiment analysis has become a popular research topic. Using an SVM classifier, Abdul-Mageed et al. [18] investigated the effect of subjectivity and sentiment classification at the sentence level for the Modern Standard Arabic language (MSA). Shoukry and Rafea [19] used 1000 tweets to apply SVM and Nave Bayes (NB) at the sentence level for sentiment classification. For sentiment analysis, Abdulla et al. [20] compared lexicon-based and corpus-based approaches.

The lexicon construction challenges and sentence analysis were addressed by Abdulla et al. [21]. Badaro et al. [22] used an English-based relating to a WordNet and the lexicon approach to create a large Arabic sentiment lexicon. Duwairi et al. [23] used crowdsourcing to collect over 300,000 Arabic tweets and label over 25,000 of them. For Arabic sentiment classification, Al Sallab et al. [24] applied three deep learning methods. Ibrahim et al. [25] used different types of text data, such as tweets and product reviews, to show sentiment classification for Egyptian dialect. The use of pre-trained models with CNN for Arabic word representation improved sentiment analysis performance, according to Dahou et al. [26].

Researchers developed aBidirectional LSTM Network (BiLSTM) with efficient feature extraction capability in [27]. The backward and forward dependencies are used to extract information from feature sequences. For evaluating the models' performance, a number of experiments was conducted on six benchmark datasets of sentiment analysis. The results show that the proposed model outperformed the other models, including the Support Vector Machine (SVM), Random Forest (RF), and LSTM. The authors of [28] investigated various deep learning (DL) models for sentiment analysis of Arabic microblogs based on LSTM and CNN. They used datasets from continuous bag-of-words (CBOW), skip-gram (SG), ASTD, and Ar-Twitter in their experiments. The experimental results revealed that LSTM outperforms CNN. In another work, the authors [29] used NB and SVM with different schemes for weighting such as n-gram sizes and TF-IDF to analyze the Arabic sentiment. They performed the experiments on AJGT dataset and they found the best performance is for the scenario of SVM classifier. The hybrid models-based DL algorithms for sentiment classification were proposed by the authors in [30]. They compared and evaluated the hybrid model to DT, RF, CNN, and RNN-LSTM, using over one million tweets from various domains. The hybrid model performed the best, according to the results.

To predict Arabic sentiment analysis, A. M. Alayba et al. [31] used a variety of machine learning models, including NB, SVM, Ridge Classifier (RDG), LR with Stochastic Gradient Descent (SGD), and the DL models, such as CNN and other methods of feature extraction. The Arabic Health Services Dataset was used to conduct the tests. Mohamed Fawzy et al. [32] stated a diversity of deep learning network models for classifying Arabic sentiment, as well as word embedding techniques. To conduct experiments, we used RNN, CNN, and Bidirectional Multi-Layer-LSTM with various word embedding. Large-scale Arabic book reviews were used in the experiments (LABR). The results revealed that the Bidirectional Multi-Layer LSTM has a high level of precision. In [33], the authors used NB, DT, LR, SVM, and DL models on Arabic tweets dataset for Saudi dialect sentiment analysis. Deep learning and SVM classifiers outperform all others in terms of accuracy. To predict sentiment analysis, some studies

used an ensemble learning (EL) approach. To predict sentiment analysis, Al-Hashedi et al. [34] used NB, RF, voting ensemble method, SGD, and LR. The author gathered COVID-19 Arabic tweets and classified them as negative or positive. The voting classifier performs well, as evidenced by the results. Alharbi et al. [35] proposed a DeepASA architecture model that included hidden layers, as well as input and output layers. Two types of deep neural networks (LSTM and GRU), and a voting classifier were used to improve the prediction performance of the model. Large Scale Arabic Hotel Reviews (HTL), Library Book Reviews Dataset (LABR), Product Reviews (PROD), Restaurant Reviews (RES), ASTD datasets, and ArTwitter were used in the experiments. The DeepASA-based approach performs well, as evidenced by the results. On the ASTD dataset of sentiment analysis for Arabic text, Oussous et al. [36] used a voting method built on top of three classifiers: NB, SVM, and Maximum Entropy. The results show that the vote algorithm is extremely accurate.

For classifying the Arabic text sentiment, Al-Saqqa et al. [37] offered an ensemble approach combines three ML classifiers, SVM, KNN, and NB using a majority voting algorithm. The datasets such as ArTwitter, Movie reviews, and a large-scale Arabic sentiment analysis were used in this study. The results of the experiments revealed that the ensemble of classifiers outperforms individual classifiers. On the Arabic sentiment dataset, Al-Azani et al. [38] studied the performance of various ensemble learning methods to enhance the performance of single classifiers, including boosting, bagging, stacking, voting, and RF. The stacking ensemble performs well, as evidenced by the results. Other researchers used ensemble learning techniques to analyze sentiment in other languages than Arabic language. For example, Sitaula et al. [39] created the NepCOV19Tweets, a Nepali Twitter sentiment dataset that was labeled negative, neutral, and positive. The authors developed some feature extraction methods based on fastText, domain-agnostic, and domain-specific feature selection techniques. Each feature selection method was implemented using different CNN models. Then, for capturing multi-scale information in order to obtain better classification results, they proposed a CNN ensemble model. The authors suggested a multi-channel CNNs (MCNNs) classification system for classifying the tweets in NepCOV19 dataset into negative, neutral, and positive sentiment [40]. A hybrid feature extraction method has been proposed for extracting syntactic and semantic features to train their proposed MCNNs model. When compared to single methods of feature extraction, the hybrid features achieved the highest accuracy result, and the MCNNs model obtained a best performance threshold. Ahmed Mohamed [41] applied Naive Bayes and SVM algorithms for Arabic sentiment analysis and El-Masri et al. [42] presented a novel method for sentiment analysis to Arabic language tweets that uses a mix of parameters related n-grams-based features and preprocessing methods. The tool analyzes the most recent tweets about the issue to determine the polarity (positive, negative, and neutral) and then displays the findings. The results of the study demonstrated that the Naive Bayes method is the most effective in detecting topic polarity. Expert and intermediate users can choose the most effective combinations of parameters for sentiment analysis with the aid of the tool. However, the related studies have a limitation in selecting the best values to hyper parameters of Naive Bayes method.

Among different ML classifiers used in previous studies, MNNB classifier is widely used for documents classification due to its ability for performing statistical analysis of text contents. It significantly simplifies textual-data classification and offers an alternative to heavy ML-based semantic analysis methods. However, the MNNB classifier has a number of hyper-parameters affects the classification task of texts and controls the decision boundary of the model itself. In this paper, an optimized MNNB classifier-based approach is proposed for improving Arabic sentiment analysis.

III. Research Methodology

This section describes the research methodology of proposed approach for social media Arabic tweets analytics. It is based on an optimized MNNB classifier model. The steps of the research methodology are shown in Fig. 1.

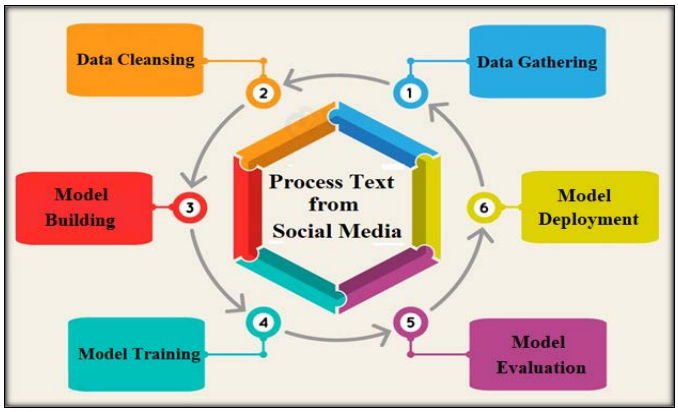

Fig. 1. Research Methodology.

A. Data Gathering

As mentioned in the scope of this research work, the model should be able to analyze Arabic sentiment from users' tweets in the highly used social networks, Twitter. The data gathered are the tweets and their retweets. Then, the data collected from the Twitter source will be stored in a dataset for training after manually labeling or for testing in the evaluation phase. Once gathering a list of Arabic topics or keywords, you can find them on Twitter and then save their related information in the storage device for further processing. The Twitter API makes it simple to search for users and returns results in JSON format. This format is easy to parse in a Python script. Dealing with social media accounts may face one complication, which is the fake accounts with similar or identical names, making them difficult to detect. Fortunately, each user object in Twitter includes a handy data field that indicates whether the account is verified, which I checked before saving the handle. The next step was to use Twitter's API to download the user's tweets and save them into a dataset once a topic or required keyword was linked to a Twitter handle.

B. Data Cleansing

It is a part of NLP since the data received contains incomplete, incorrect, inaccurate or irrelevant parts of the data records that required replacing, modifying, or deleting them.

Therefore, it will be used in the proposed methodology for English and Arabic text. It can perform the following:

- Delete the empty records.
- Delete the retweets.
- Remove the Hashtags, pictures, and links.
- Remove usernames from mentions.
- Remove English and Arabic stop words.
- Normalize the words.
- Stemming the words.

After performing the previous functions, the data will be ready to be used in the model for training and testing purposes.

C. Model Building

After data cleansing, the selection of a model is the next step in the research methodology process. Over the years, researchers and data scientists have developed a variety of models. Some are better suited to image data, while others are better suited to sequences (such as voice or text), numerical data, and text-based data. A multinomial naïve Bayes (MNNB) classifier is a simple and effective ML model to deal with text features. The MNNB classifier is often used as a starting point for sentiment analysis.

The core idea behind the MNNB technique is to use the joint probabilities of words and classes to find the probabilities of classes given to texts.

Given a vector of dependent features $(f_1, \dots, f_n)$ and the class L_k, Bayes' theorem can be expressed mathematically as:

$$P(L_k|f_1, \dots, f_n) = \frac{P(L_k)P(f_1,\dots,f_n|L_k)}{P(f_1,\dots,f_n)} \quad (1)$$

For the given class L_k and consistent with the assumptions of naïve conditional independence, each feature f_i is conditionally independent of every other feature f_j where $i \neq j$.

$$P(f_i|L_k, f_1, \dots, f_n) = P(f_i|L_k) \quad (2)$$

Thus, it can be simplified the relation to be as:

$$P(L_k|f_1, \dots, f_n) = \frac{P(L_k)\prod_{i=1}^{n} P(f_i|L_k)}{P(f_1,\dots,f_n)} \quad (3)$$

Because $P(f_1, \dots, f_n)$ is constant and if the feature values of the known variables, the following rule for classification can be employed:

$$P(L_k|f_1, \dots, f_n) \propto P(L_k)\prod_{i=1}^{n} P(f_i|L_k)$$

$$\Downarrow$$

$$\hat{y} = arg\max_k P(L_k)\prod_{i=1}^{n} P(f_i|L_k) \quad (4)$$

Log probabilities can be used to avoid underflow.

$$\hat{y} = arg\max_k(\ln P(L_k) + \sum_{i=1}^{n} \ln P(f_i|L_k)) \quad (5)$$

According to the distribution of $P(f_i|L_k)$, the assumptions made by the Naive Bayes classifier differs between them, whereas $P(L_k)$ is relatively defined as the frequency of class L_k in the training data set.

The distribution of multinomial naïve Bayes is parametrized by the vector $\theta_k = (\theta_{k1}, \dots, \theta_{kn})$ for each class L_k , where n represents the number of features (vocabulary size) and θ_{ki} denotes the probability $P(f_i|L_k)$ of feature i that appears in a sample that belongs to the class L_k.

The estimation of parameters θ_k can be obtained by a smoothed version of maximum likelihood (i.e., relative frequency counting) as follows:

$$\hat{\theta}_{ki} = \frac{N_{ki}+\alpha}{N_k+\alpha n} \quad (6)$$

Where N_{ki} represents the number of times feature i appears in class a sample k of the training dataset T. The total count of all features for class L_k is N. The smoothing parameter alpha (α) can have a value greater than zero and less than or equal 1 $0 < \alpha \leq 1$ for features, which are not present in the learning samples and to prevent from division by zero probabilities in further computations.

Setting $\alpha = 1$ is termed as Laplace smoothing, while $\alpha < 1$ is termed as Lidstone smoothing. Therefore, the final decision rule is written as:

$$\hat{y} = arg\max_k(\ln P(L_k) + \sum_{i=1}^{n} \ln \frac{N_{ki}+\alpha}{N_k+\alpha n}) \quad (7)$$

Unfortunately, the way to know what α gives the most accurate responses is through iterating over all values of α on the training set and this way is a hard problem. The optimized MNNB classifier used in this research determines the optimal value for hyper-parameters is through a grid search over possible parameter values and using cross-validation evaluation technique for each value. The flowchart of the optimization process for MNNB classifier-based proposed approach is shown in Fig. 2.

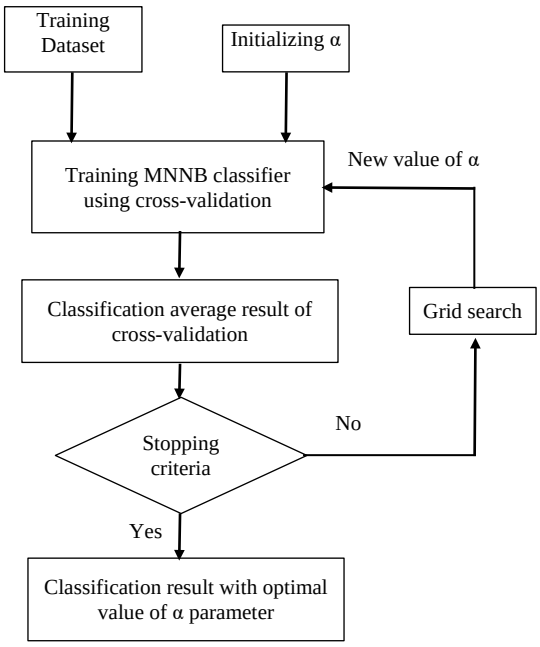

Fig. 2. Flowchart of Optimized MNNB Classifier-based Approach.

D. Model Training

The training phase is where the main process of developing the baseline and proposed models begins. Three different datasets are used for training the models of the research approach. By analyzing the characteristics of Arabic tweets, the models learn to classify their sentiments. For the MNNB classifier, the model is trained through mining the statistics from the training dataset. The utilization of frequency and likelihood tables for each feature facilitates the calculation process.

Possible feature values are grouped together in a frequency table derived from the observations. Instead of counting the number of occurrences, the likelihood table shows the probability values for each class. The multiplication and comparison operations are used to determine the class of an observation using these tables. The training process of MNNB classifier can be summarized in the following steps:

- Convert the Arabic text data into words vectors such as using TFIDF vectorizer or count vectorizer.
- Calculate the counts based on the class.
- Calculate all the likelihood probabilities.
- Calculate the prior probability.
- Calculate the posterior probability.

Feature vectors represent the frequency with which specific events were generated by a multinomial distribution. This is the most commonly used event model for Arabic text classification. This algorithm is used to solve problems with Arabic text classification. This method could be used to determine whether a tweet belongs in the ‘positive’ or ‘negative’ category, for example. It takes advantage of the current words’ frequency as a feature.

E. Model Evaluation

The model must be tested after it has been trained with train data. The goal of testing is to see how the model performs in real-world situations. During this phase, we can assess the model's accuracy. In our case, the model uses the learning from the previous phase to try to identify the type of fruit. The evaluation phase is crucial, as it allows us to see if the model achieves the goal we set for it. If the model does not perform as expected during the testing phase, the previous steps must be repeated until the required accuracy is achieved. As stated, it should not use the same data that was used in the training phase. For evaluation, it should have to use the separate data splitter from the dataset.

The only thing that classification models care about is whether or not the result is correct. When making classification predictions, such as the one we used, there are four possible outcomes. True negatives, true positives, false negatives, and false positives are the four types. On a confusion matrix, these four outcomes are plotted. After making predictions on the test data, you can create the matrix and categorize each prediction as one of the possible outcomes. The percentage of correct predictions made by the test data determines the model's accuracy. The model's accuracy can be determined by dividing the number of correct predictions by the total number of predictions. Classification models are also evaluated using other metrics such as accuracy.

F. Model Deployment

Model deployment is the integrating process of ML model in an existing construction environment to make decisions for data-driven business. It is the last step in the ML process and one of the most time consuming steps. Traditional model-building languages are frequently incompatible with an organization's IT systems, which force programmers and data scientists to spend brainpower and valuable time for rewriting them.

A model must be successfully deployed into production before it can be used for practical decision-making. If you can't rely on your model to provide practical insights, then its impact is severely limited. One of the difficult aspects of achieving value from ML is model deployment. IT teams, software developers, data scientists, and business professionals must work together to ensure that the model works reliably in the organization's production environment. This is a significant challenge due to there is frequently a mismatch between the programming-language used to create an ML model and the languages that the production system understands. Model re-coding can add weeks or months to the project timeline.

The deployment of ML models is the final step in the process. ML models are typically developed and tested using training and testing datasets in a local or offline environment. When a model is deployed, it is placed in a live environment and is exposed to new and unknown data. As the model performs the task, it was trained for working on live data, the model begins to bring to the organization a return on investment.

Containerization is becoming increasingly popular as a tool for deploying ML models. Containers are a common environment to deploy the models because they simplify updating and deploying different parts of the model. Containers are intrinsically scalable, as well as able to provide a consistent environment for the model function. Kubernetes and other open-source platforms are utilized for managing and automating the container management aspects such as scaling and scheduling.

IV. Experimental Results and Discussions

Before presenting the experimental results, this section starts by presenting the number of instances for each class in the training and test datasets. The adopted ML models are trained on three different datasets. Each dataset has a larger number of tweets as shown in Fig. 3 to 5. For the dataset 1, Fig. 3 illustrates that the number of negative instances is 976, and the positive instances is 1046, as well as the number of natural instances is 724. For the dataset 2, Fig. 4 demonstrates that the number of negative instances is 2588, and the positive instances are 1817, and the number of natural instances is 1587. Similarly, the dataset 3 contains 228 negative instances, 263 positive instances, and 192 natural instances as seen in Fig. 5. The test dataset is also used to evaluate the trained models for sentiment classification of Arabic tweets. Fig. 6 displays the number of instances in the test dataset.

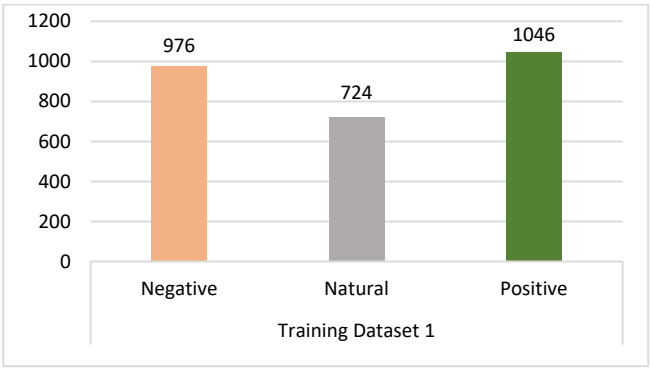

Fig. 3. Number of Instances in Training Dataset 1.

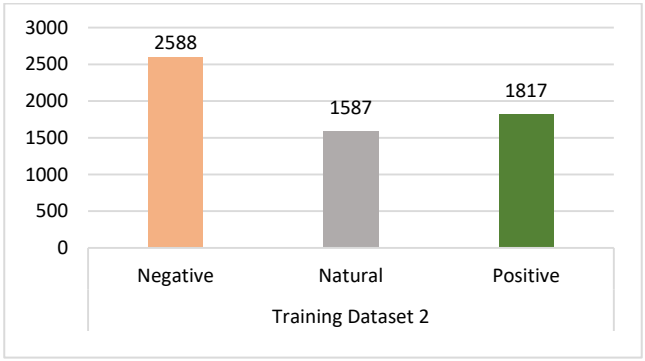

Fig. 4. Number of Instances in Training Dataset 2.

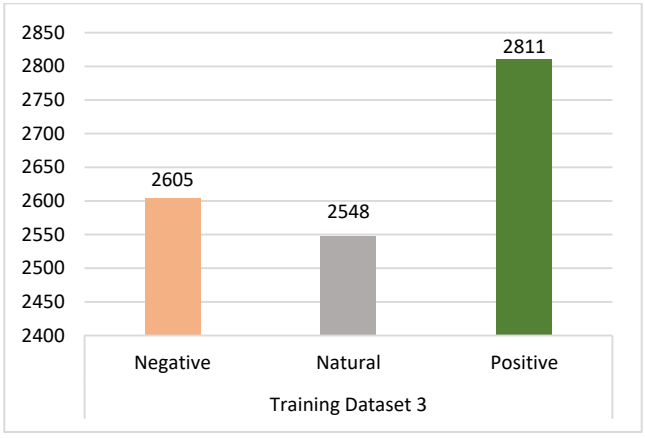

Fig. 5. Number of Instances in Training Dataset 3.

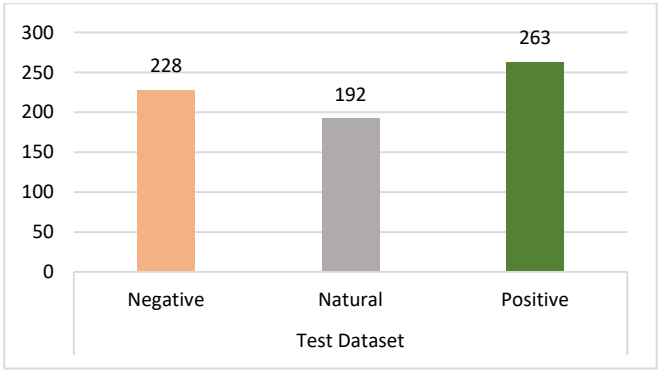

Fig. 6. Number of Instances in Test Dataset.

From Table I to Table VI, the confusion matrices of test set classification for the optimized MNNB trained on the three datasets is given using TFIDF and count vectorizers. Table VII and Table VIII show the evaluation results, which show each classifier's performance on positive, natural and negative instances, as well as their overall performance.

TABLE I. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 1 AND USING TFIDF VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	139	15	74
Natural	71	37	84
Positive	44	19	200

TABLE II. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 1 AND USING COUNT VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	136	25	67
Natural	62	57	73
Positive	49	23	191

TABLE III. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 2 AND USING TFIDF VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	167	12	49
Natural	109	28	55
Positive	73	19	171

TABLE IV. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 2 AND USING COUNT VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	136	26	66
Natural	77	50	65
Positive	53	30	180

TABLE V. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 3 AND USING TFIDF VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	111	39	78
Natural	72	46	74
Positive	33	36	194

TABLE VI. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 3 AND USING COUNT VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	115	42	71
Natural	69	54	69
Positive	34	36	193

TABLE VII. CLASSIFICATION ACCURACY RESULT USING TFIDF VECTORIZER

Classifier	Training Dataset 1	Training Dataset 2	Training Dataset 3
SVM	0.511	0.511	0.515
SVM Linear Kernel	0.530	0.531	0.518
RF	0.488	0.486	0.463
GaussianNB	0.483	0.474	0.466
MNNB	0.539	0.521	0.514
Optimized MNNB	**0.551**	**0.536**	**0.514**

TABLE VIII. CLASSIFICATION ACCURACY RESULT USING COUNT VECTORIZER

Classifier	Training Dataset 1	Training Dataset 2	Training Dataset 3
SVM	0.492	0.476	0.460
SVM Linear Kernel	0.526	0.518	0.508
RF	0.488	0.502	0.466
GaussianNB	0.482	0.480	0.473
MNNB	0.546	0.542	0.530
Optimized MNNB	**0.562**	**0.542**	**0.530**

From the above results, the following observations are made:

- From the three datasets used for training as shown in Fig. 3 to 5, we can see that there is a diversity in the number of instances for each class to see the effect of class size for training the models.
- From Table I to Table VI, the confusion matrices show that the number of instances, which are correctly classified for the optimized MNNB classifier is improved by selecting the optimal values of α hyper-parameter, especially on neutral and negative instances.
- As shown in Table VII and Table VIII, the optimized MNNB classifier obtains high accuracy results by using count vectorizer representation for Arabic tweets as features rather than TFIDF vectorizer.
- SVM with linear kernel works well than other models, especially using TFIDF vectorizer but not better than MNNB and optimized MNNB classifiers.
- For sentiment classification of Arabic tweets, MNNB model is preferable as a generative model. It outperforms the other baseline classifiers.

V. CONCLUSION AND FUTURE WORK

Arabic sentiment analysis using machine learning methods has become an important research task for developing various applications. In this paper, an optimized MNNB classifier-based approach is presented for improving Arabic sentiment analysis. It aims to select the optimal value of the MNNB's alpha hyper-parameter to control the decision boundary of the model itself. The sentiment classification experiments of the research are conducted using a large-scale data sets. The baseline and optimized MNNB classifiers are trained on three datasets and tested on a different separated test set to show the performance of developed approach. The experimental results on the test set revealed that the optimized MNNB classifier of proposed approach outperforms the traditional MNNB classifier and other baseline classifiers. The accuracy rate of the optimization approach is increased by 1.6% compared with using the default values of the classifier's hyper-parameters. The output from the study shows that a MNNB classifier with count vectorizer as features can achieve a high performance compared to the other baseline classifiers. Because there are a large number of Arabic tweets features that are likely to be noisy, a feature selection scheme can be investigated in future work. Previous classification studies have shown that feature selection is critical for classification task success. Another promising extension of this research would be to classify Arabic tweets on different scales instead of just positive, neutral, and negative classes. Moreover, a combination of different classifiers and deep learning methods will be explored.

ACKNOWLEDGMENT

"The author is thankful to the Deanship of Scientific Research, College of Computer and Information Sciences (CCIS) at King Saud University for funding this research."

REFERENCES

[1] M. Birjali, M. Kasri and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," Knowledge-Based Systems, vol. 226, p. 107134, 2021.

[2] D. Antonakaki, P. Fragopoulou and S. Ioannidis, "A survey of twitter research: Data model, graph structure, sentiment analysis and attacks," Expert Systems with Applications, vol. 164, p. 114006, 2021.

[3] S. Zad, M. Heidari, J. H. Jones and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0285-0291: IEEE.

[4] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari and A. Hilal, "Preprocessing arabic text on social media," Heliyon, vol. 7, no. 2, p. e06191, 2021.

[5] R. Bensoltane and T. Zaki, "Aspect-based sentiment analysis: An overview in the use of arabic language," Artificial Intelligence Review, pp. 1-39, 2022.

[6] M. Hijjawi and Y. Elsheikh, "Arabic language challenges in text based conversational agents compared to the english language," International Journal of Computer Science Information Technology, vol. 7, no. 5, pp. 1-13, 2015.

[7] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali and I. Abunadi, "Arabic natural language processing and machine learning-based systems," IEEE Access, vol. 7, pp. 7011-7020, 2018.

[8] B. Liu, "Sentiment analysis and subjectivity," Handbook of natural language processing, vol. 2, no. 2010, pp. 627-666, 2010.

[9] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: A survey of techniques," arXiv preprint arXiv:.06971, 2016.

[10] L.-w. Ku, Y.-s. Lo and H.-h. Chen, "Using polarity scores of words for sentence-level opinion extraction," in Proceedings of NTCIR-6 workshop meeting, 2007, pp. 316-322: Citeseer.

[11] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of html documents," in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 1075-1083.

[12] L. Yang, B. Liu, H. Lin and Y. Lin, "Combining local and global information for product feature extraction in opinion documents," Information Processing Letters, vol. 116, no. 10, pp. 623-627, 2016.

[13] L. Yang, H. Lin, Y. Lin and S. Liu, "Detection and extraction of hot topics on chinese microblogs," Cognitive Computation, vol. 8, no. 4, pp. 577-586, 2016.

[14] W. Han, Z. Tian, Z. Huang, S. Li and Y. Jia, "Topic representation model based on microblogging behavior analysis," World Wide Web, vol. 23, no. 6, pp. 3083-3097, 2020.

[15] L. Huang, S. Li and G. Zhou, "Emotion corpus construction on microblog text," in Workshop on Chinese Lexical Semantics, 2015, pp. 204-212: Springer.

[16] P. Tiwari, P. Yadav, S. Kumar, B. K. Mishra, G. N. Nguyen et al., "Sentiment analysis for airlines services based on twitter dataset," Social Network Analytics: Computational Research Methods & Techniques, vol. 149, 2018.

[17] L. Yang, S. Zhang, H. Lin and X. Wei, "Incorporating sample filtering into subject-based ensemble model for cross-domain sentiment classification," in Chinese computational linguistics and natural language processing based on naturally annotated big data: Springer, 2015, pp. 116-127.

[18] M. Abdul-Mageed, M. Diab and M. Korayem, "Subjectivity and sentiment analysis of modern standard arabic," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 587-591.

[19] A. Shoukry and A. Rafea, "Sentence-level arabic sentiment analysis," in 2012 international conference on collaboration technologies and systems (CTS), 2012, pp. 546-550: IEEE.

[20] N. A. Abdulla, N. A. Ahmed, M. A. Shehab and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), 2013, pp. 1-6: IEEE.

[21] N. Abdulla, S. Mohammed, M. Al-Ayyoub and M. Al-Kabi, "Automatic lexicon construction for arabic sentiment analysis," in 2014 International Conference on Future Internet of Things and Cloud, 2014, pp. 547-552: IEEE.

[22] G. Badaro, R. Baly, H. Hajj, N. Habash and W. El-Hajj, "A large scale arabic sentiment lexicon for arabic opinion mining," in Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP), 2014, pp. 165-173.

[23] R. M. Duwairi, R. Marji, N. Sha'ban and S. Rushaidat, "Sentiment analysis in arabic tweets," in 2014 5th international conference on information and communication systems (ICICS), 2014, pp. 1-6: IEEE.

[24] A. Al Sallab, H. Hajj, G. Badaro, R. Baly, W. El-Hajj et al., "Deep learning models for sentiment analysis in arabic," in Proceedings of the second workshop on Arabic natural language processing, 2015, pp. 9-17.

[25] H. S. Ibrahim, S. M. Abdou and M. Gheith, "Sentiment analysis for modern standard arabic and colloquial," arXiv preprint arXiv:.03105, 2015.

[26] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud and P. Duan, "Word embeddings and convolutional neural network for arabic sentiment classification," in Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers, 2016, pp. 2418-2427.

[27] H. Elfaik, "Deep bidirectional lstm network learning-based sentiment analysis for arabic text," Journal of Intelligent Systems, vol. 30, no. 1, pp. 395-412, 2021.

[28] S. Al-Azani and E.-S. M. El-Alfy, "Hybrid deep learning for sentiment polarity determination of arabic microblogs," in International Conference on Neural Information Processing, 2017, pp. 491-500: Springer.

[29] K. M. Alomari, H. M. ElSherif and K. Shaalan, "Arabic tweets sentimental analysis using machine learning," in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2017, pp. 602-610: Springer.

[30] M. H. Abd El-Jawad, R. Hodhod and Y. M. Omar, "Sentiment analysis of social media networks using machine learning," in 2018 14th international computer engineering conference (ICENCO), 2018, pp. 174-176: IEEE.

[31] A. M. Alayba, V. Palade, M. England and R. Iqbal, "Improving sentiment analysis in arabic using word representation," in 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), 2018, pp. 13-18: IEEE.

[32] M. Fawzy, M. W. Fakhr and M. A. Rizka, "Word embeddings and neural network architectures for arabic sentiment analysis," in 2020 16th International Computer Engineering Conference (ICENCO), 2020, pp. 92-96: IEEE.

[33] M. E. M. Abo, N. Idris, R. Mahmud, A. Qazi, I. A. T. Hashem et al., "A multi-criteria approach for arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection," Sustainability, vol. 13, no. 18, p. 10018, 2021.

[34] A. Al-Hashedi, B. Al-Fuhaidi, A. M. Mohsen, Y. Ali, H. A. Gamal Al-Kaf et al., "Ensemble classifiers for arabic sentiment analysis of social network (twitter data) towards covid-19-related conspiracy theories," Applied Computational Intelligence Soft Computing, vol. 2022, 2022.

[35] A. Alharbi, M. Kalkatawi and M. Taileb, "Arabic sentiment analysis using deep learning and ensemble methods," Arabian Journal for Science Engineering, vol. 46, no. 9, pp. 8913-8923, 2021.

[36] A. Oussous, A. A. Lahcen and S. Belfkih, "Impact of text pre-processing and ensemble learning on arabic sentiment analysis," in Proceedings of the 2nd International conference on networking, information systems & security, 2019, pp. 1-9.

[37] S. Al-Saqqa, N. Obeid and A. Awajan, "Sentiment analysis for arabic text using ensemble learning," in 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), 2018, pp. 1-7: IEEE.

[38] S. Al-Azani and E.-S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text," Procedia Computer Science, vol. 109, pp. 359-366, 2017.

[39] C. Sitaula, A. Basnet, A. Mainali and T. B. Shahi, "Deep learning-based methods for sentiment analysis on nepali covid-19-related tweets," Computational Intelligence and Neuroscience, vol. 2021, 2021.

[40] C. Sitaula and T. B. Shahi, "Multi-channel cnn to classify nepali covid-19 related tweets using hybrid features," arXiv preprint arXiv:.10286, 2022.

[41] A. Mohamed, "Svm and naive bayes for sentiment analysis in arabic," PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1631367/v1], 2022.

[42] M. El-Masri, N. Altrabsheh, H. Mansour and A. Ramsay, "A web-based tool for arabic sentiment analysis," Procedia Computer Science, vol. 117, pp. 38-45, 2017.

AUTHORS' PROFILE

Dr. Ahmed Alsanad, is an Associate Professor of Information System Department and chair member of Pervasive and Mobile Computing, CCIS, at the King Saud University, Riyadh, KSA. He received his Ph.D. degree in Computer Science from De Montfort University, Unit Kingdom in 2013. His research interests include Cloud Computing, Health Informatics, ERP and CRM. He has authored and co-authored more than 12 publications including refereed IEEE/ACM/Springer journals, conference papers, and book chapters.