JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

Political Post Classification based on Firefly and XG Boost

Ahmed Assim Nsaif[1], Dhafar Hamed Abd[1,2]
[1]Informatics institute for postgraduate student, Baghdad, Iraq
[2]Department of Computer Science Al-Maarif University College, Anbar, Iraq

Ahmedalsray50@gmail.com, Dhafar.dhafar@gmail.com

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10

Abstract:

Opinion mining is used in practically every aspect of human life and has a substantial effect in our behaviors. There is a lot of data that displays people' opinions in many domains, like business and politics, due to the expansion and use of online technology. To generate our vector, we used the firefly method to select the finest words from political Arabic posts and looked into two feature extractions: term frequency and term frequency inverse document frequency. These features utilized with XGBoost algorithm to classified the right class into (Revolutionary, Conservative, andReform). Accuracy, F1-score, recall,precision,and number of correct predict were calculated to measure the applied classifiers' performance.The results expose that the TF conclude the best results in accuracy of 98.052% with length of features 210.

Keywords: XGBoost; firefly algorithm; feature selection; political Arabic post; Term Frequency; Term Frequency-Inverse Document Frequency

1. Introduction

In recent years, the use of web resources such as online review sites, social networking sites, personal blogs, and other similar sites has increased, allowing participants to express or give their opinions, ideas, and remarks on a variety of topics. It's critical to collect and analyse these remarks in real-life scenarios. For example, before purchasing a service or product, every customer would like to hear what other customers have to say. Likewise, a corporation wants to know what the consumer thinks so that they may enhance and modify their product to meet his needs. A political party's goal in the political realm is to forecast voter trends and preferences.

As the subjective textsnumber in forums, blogs, and social media has increased since the 2000s, several researchers have employed a text classification techniques like sentiment classification[1]. opinion extraction, review mining, opinion mining, sentiment analysis, and subjectivity analysis were all terminology used by a few researchers to describe the sentiment categorization[2].

Opinion polls, e-commerce, and education have all profited from sentiment analysis[3]. The company analyses customer evaluations and watches social media to identify people's attitudes of their products and services and take relevant action in a timely manner. According to some researchers, stock prices and social media sentiment analyses are linked, and future stock prices can be forecasted utilizing microblogs' sentiments such as Twitter[4].

The purpose of this paper is to look into two topics: the general political regimes classification and how to classify the Arabic dataset. In terms of the latter, political regime labelling is impossible to carry out without a clear description of these regimes.
Many sectors associated with the study of these resources to extract relevant information have seen a large increase in the prominence of newspapers, ideological websites, and online social networks. Sentiment analysis (SA), also referred to as sentiment orientation, or opinion mining is the process of determining the orientation of unstructured data. It is a categorization exercise in which the sentiment or point of view expressed in an article or sentence is classified as neutral, positive, or negative. Because the major of studies are in English, and some are written in other languages, conducting such research in Arabic is critical[5]. The Arabic language, on the other hand, is distinct from English and has its own set of challenges and issues.

Arabic is regarded as the world's fifth most widely spoken language. According to the most recent statistics, over 422 million people speak Arabic as their first language, with 250 million speaking it as a second. There are 28 letters in the Arabic alphabet, and there is no lower or upper case in Arabic letters. Arabic is written in a right-to-left orientation[6]. In comparison to the profusion of studies on English opinions, emotions, attitudes, and sentiments, only a few studies have been undertaken on Arabic language. Previous studies focused on specific items, such as political pieces, which were classified into the political class, and sports articles, which were classified into the sport class. This study tries to employ polarity orientation recognition, and collect data on Arabic political articles from newspapers, ideological websites, and social media.

Documents, sentences, and aspects are the three primary levels of classification in SA. This paper seeks to classify an opinion piece as expressing a negative or positive opinion at the document level. The entire article is regarded as a unit of basic data (talking about one topic). The goal of this research is to examine publications in order to establish their political leanings. The choice was based on a crude set theory and the sentiment language of the texts under consideration. Revolutionary, Conservative, and Reformist are the three political orientations we consider in this study.

Selecting features can be done in one of three ways: "embedded method, filter, and wrapper" Filter methods, which act as pre-processing for scoring the features, are used to choose and supply the highest rated features to a predictor. Wrapper approaches employ the predictor's performance as a feature selection criterion;thus, the predictor is wrapped around a search algorithm that discovers the subset with the best predictor performance. Embedded methods combine the advantages of both the wrapper and filter approaches, incorporating variable selection as step of the procedure rather than dividing it into train and test sets. This research adopts the technique of wrapping, which selects the best subset of words to generate the vector using the firefly algorithm.

2. Related works

Previous researches in this field can be classified on the base of the utilizedtechnique that achieve the sentiment classification. To procure text's sentiment information, the machine learning approach primarily uses

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

linguistic models. This experiment was conducted to understand more about the machine learning approach.

S. Alterkavi et al.[7], proposed a new approach for author verification in OSNs, which involved some twitter's textual features extraction and using the algorithm of XGBoost as an efficient pre-processing tool to find useful propertiessubsets. Methods that used are LSR, SVM, RF, XGBoost. Obtained datasets from Twitter with size of 16124. The pre-processing approaches are Retweet deleting, Tokenization, Remove (hash tag, punctuation, percentage, Month, emoji's). The best results (Metrics of CADx) ACC is 90.49%, f1 is 94%, Precision is 90.2% and recall is 96.7%. The author's difficulties with feature selection have been investigated by treating it like an MCDM problem. Many parameters, such as booster type, can be tuned to improve results, according to the XGBoost algorithm description. The avoiding over-fitting parameters and learning process parameters are related to booster parameters. In addition, there still are tree related parameters such as Max Depth, which represents the maximum tree depth value.

AlZoubi et al.[8], resolve one of the state-of-the-art models established to address the problem of the research emotion intensity analysis of Arabic tweets to determine the polarity (neutral, negative, or positive) for a given Tweet. Method is used BiGRU_CNN, CNN, XGB, Ensemble, and SVM. Obtained datasets from tweeter where the collected tweets were gathered utilizing the twitter API on the base of somewords of emotion with size of5601. The pre-processing approaches are Tokenization, normalization, remove elongation and numbers and non-Arabic characters were also removed. The best results (Metrics of CADx) are ACC is 68.4%. Challenges of author are the tweets analyzing as adata source of social media. It is written in a slang term that could contain grammatical faults, making it difficult to interpret by machines. The plan is to look at the usage of transformers like ULM-FiT, ELMo, and BERT in measuring Arabic tweets' emotion intensity.

A. N. Azhar and M. L. Khodra [9]will examine how to convert the ABSA problem from a single sentence to a sentence-pair classification task by building auxiliary sentences basing on Sun, et al. CNN-XGBoost was the method utilized. Datasets with a total size of 9448 were obtained from a former Indonesian accommodation network orchestrator. Companies have commonly employed sentiment analysis as a preprocessing approach to automatically collect opinions about their services or products. The best score for f1 (CADx metrics) is 89.58. The difficulties are numerous. Out-of-vocabulary (OOV) words continue to cause a lot of misclassifications. As a result, new techniques must be used to create a better model. It is suggested that diverse forms of text, such asacademic papers, news, and reviews be combined for pretraining BERT so that the model is exposed to a variety of contexts and text structures, resulting in better representation of language.

M. S. Bhatti et al.[10], established a significant landmark in Urdu news classification by finding cosine similarity news. FastText is used by Urdu among to extract the word's semantic meaning. It also specifies how to extract document's features and select sparsity reduction features in order to improve the efficiency of machine learning algorithms. XGBoost, Bagging, and Nave Bayes are the methods employed. Datasets having a total size of 20105 were painstakingly retrieved from reliable Pakistani Urdu news sources. The pre-

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

processing approaches are removing (duplicate, null values, punctuation, stop-words) tokenizing sentences. The best results (Metrics of CADx) f1 is 85%, Precision is 88% and recall is 84%. Challenges are The Urdu language has not had much research done on it, and there are no available readily datasets, there are not many publicly available datasets, The text classification used for other languages that are comparable to Urdu cannot be applied since there is no proper delimitation between words in Urdu. It is suggested that a large collection of Urdu news documents be gathered and made available for future studies.

M. Bobur, et al.[11], aims to overcome the problem of detecting irregularities in court acts. Methods used are XGBoost and LR. The datasets are private data with size of 400000. The pre-processing approaches are various filtering methods. The best results (Metrics of CADx) ACC is 75.44%. The current outlier methods of search use high dimensional domains where data can have hundreds of dimensions, which is a problem. This method takes a lot of time and money and is not very effective. The suggestions are plans for making the model better: to improve the speed and accuracy of processing one request.

D. Endalie et al.[12], provide a model of deep learning for the classification of Amharic news documents. It intends to enhance the categorization accuracy of Amharic documents through the use of deep learning and quick text pretraining. Methods that used are CNN, RF, XGB, SVM, MLP, and DT. Obtained datasets from Amharic news documents from sixmain news classifications with 3600 sizes. The pre-processing approaches are Normalization, Stemming, and Tokenization. The experimental result showed that best results (Metrics of CADx) ACC is 87.58%. Obstacles In machine learning, reducing the high-dimensional feature space dimensionality which is among the most complicated problems. It is exceedingly difficult to calculate the filter's number and the best filter size for a certain task.

D. Godavarthi and M. S. A[13], suggested a machine learning based approach for extracting the scientific literature's information using text categorization applications of natural language processing. The primary objective is to classify the covid-21 abstracts with their corresponding journals so that a researcher can refer to his required journal's papers rather than browsing through all the articles. Methods that used are KNN, MLP and XGB. Obtained datasets from CORD-19 with size of 200000. The pre-processing approaches are Data cleaning, Tokenization, stemming, remove stop word and lemmatization. The experimental result showed that best results (Metrics of CADx) ACC is 84%. Finding solutions to problems about COVID-19 has grown to be a significant difficulty for the medical community as a result of the growing literature. We have put forth a system based on machine learning that mines the scientific literature for data using text categorization applications of NLP. It is necessary to create a deep-learning system that is more accurate in the future.

R. H. Hama Aziz et al. [14], proposes a new ensemble classifier approach that combines various sets of features with ensemble classification by integrating various base classifiers, which are slow learners, into different classifiers. This study aims to find the social media of sentiments polarity and put them into negative or positive groups. Methods used are NB, KNN, LR, RF, DT, SGD, AdaBoost, XGB, Bagging, SentiXGboost. Obtained datasets from Twitter with size of 10551. The pre-processing approaches are Tokenization, removing stop words and stemming. The experimental result showed that

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

best results (Metrics of CADx) are ACC is 90.8%, f1 is 94%, Precision is 92.7%and recall is 98.1%. The primary Sentiment Analysis obstacles, particularly in social media, include (1) extensive but inconsistent and ad hoc usage of abbreviations and acronyms, (2) informal language, and (2) the message's brevity. There will be a need to enhance the effectiveness of sentiment classifiers that self-contained which can be utilized on data from various disciplines, as advised.

M. I. Hossain, et al.[15], suggested to use XGBoost, the Random Forest Classifier, and for rating classification, a Logistic Regression technique with TF-IDF Vectorizer is utilized. According to our knowledge, no prior research has been conducted on the "GrammarandProductReviews" dataset, which is the impetus for this study. Therefore, the selection of the dataset for this study was motivated by the desire to analyze the performance of a novel dataset using several algorithms of machine learning. Methods that used are RF, LR and XGBoost. Obtained datasets from Kaggle with size of 71045. The pre-processing approaches are Remove (punctuation, stop words), Tokenization, Stemming and Lower-case conversion. The experimental result showed that best results (Metrics of CADx) ACC is 84.7 %. The difficulties are Sometimes consumers buy an online product and leave a text review, but they are averse to leaving a number rating, most frequently a star rating. However, producers want product ratings in order to analyze their business. It is advised to forecast rating based on synonyms and combine it with the suggested algorithm.

H. Karayiğit et al. [16], presents the Abusive Turkish Comments (ATC) dataset for detecting Turkish abusive comments from Instagram. The first goal of this research was to develop a dataset for detecting abusive content of messages. The second goal was to create a raccurate and obust SBTC model for detecting abusive utterances in ATC datasets. The study was motivated by the fact that image's abusive comments can be harmful and humiliating to persons who share photos. Methods used are SVM, NB, RF, LR, DT, AdaBoost and DC+BoW+TF-IDF+XGBoost. Obtained datasets from ATC with size of 30354. The experimental result showed that best results (Metrics of CADx) are Precision is 96.6%and recall is 80.4%. It is time consuming and difficult to implement a comment filter in other than English languages. It is suggested that new SBTC models be developed on the dataset ATC utilizing different attitudes (i.e., aggression, sexism, racism, and hate).

Z. M. Li, et al. [17], The purpose of this study was to identify the rumor refuter characteristics of social media. Two groups of people's Information were mined: refuters; fromlists of retweets, and stiflers; from thelists of comments. The overall objective was the number of samples determination that required to gain a stable, available F1 Score/AUC Score. The Motivating factor is decision making that support and identify the Rumor countermeasures and adapting features of user for Rumor management. Method that used is XGBoost. Obtained datasets from Weibo with size of 58807. The pre-processing approach is data cleaning. The experimental result showed that best results (Metrics of CADx) f1 is 74.90%. The difficulty is that there is no agreed-upon definition of rumors from many academic perspectives, also no established standards of evaluation for rumors. As a result, it is practically impossible to identify rumors without encountering numerous challenges and issues. It is advised to look at the characteristics of refuter in a wider variety of microblog samples, with hope of accounting for any potential bias with a lot of data. To

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

more precisely define the refuter community, we will also take into account more unique and customized features in addition to demographic characteristics. Future studies will also examine the power and influence of refuters.

3. Background theory

In this section will show you two techniques which are XGBoost algorithm were used as classification model and firefly algorithm as feature selection to select best subset from political Arabic post dataset.

3.1 XGBoost algorithm

Extreme gradient boosting or XGBoost[1] is there a well-known method of gradient boosting (ensemble) that enhances the speed and the performance of the machine learning algorithms, tree-based (sequential decision trees). Tianqi Chen was createdXGBoost and itmaintained by the Distributed (Deep) Machine Learning Community (DMLC) [2]. Because of its success with structured and tabular data, this most common algorithm is applied for machine learning in competitions. It is open source and free software. XGBoost packages were available for R and Python, but are now also accessible for Julia, Scala,Java, and etc.

XGBoost is a decision tree ensemble based on gradient boosting designed to be highly scalable[3]. XGBoost is an algorithm that, like gradient boosting, works by minimizing a loss function in order to construct the objective function's additive expansion. The loss function variation is applied in order to manage the complexity of the decision trees utilized by XGBoost. This is necessary due to the fact that the algorithm relies solely on decision trees as its base classifier.

$$L_{xgb} = \sum_{n=1}^{N} L\big(y_i, \mathrm{F}(x_i)\big) + \sum_{m=1}^{M} \Omega(h_m) \quad (1)$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda||w||^2 \quad (2)$$

where T represents the total number of leaves in the tree and w represents the scores that are generated by each leaf's output. A pre-pruning approach can be developed by incorporating this loss function into the split criterion of decision trees. When the value of is increased, the tree becomes simpler. The value of γ determines the level of minimum loss reduction gain that must be achieved before an internal node can be divided. Shrinkage is an additional regularization parameter that can be used in XGBoost[4], reducing the step size of the additive expansion. Lastly, it is also possible to limit the trees complexity by employing various techniques, such as the trees depth, etc. A tree complexity reduction benefit is the faster training of the models and using less storage.

Additionally, the techniques of randomization are included in XGBoost to decrease overfitting and increase training speed[5]. XGBoost includes two randomization techniques: random subsamples for training column subsampling and individual trees at the tree node levels and trees.

Additionally, XGBoost employs a number of techniques to increase the speed of training of decision trees that are not tied

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

to ensemble accuracy directly[6]. XGBoost's primary objective is to reduce the amount of time and effort required by the algorithm used to construct decision trees by finding the optimal split in the data. This is the part of the program that takes the most processing power[7]. In most cases, algorithms of split-finding will consider all of the possibilities and select the option that will result in the greatest gain. In order to identify the ideal split for each node, this needs a linear scanning of all of the sorted properties. XGBoost uses a compressed column-based structure, in which the data is maintained presorted, so that it does not have to constantly go through the process of sorting the data in each node[8]. Thus, each attribute sorted only once. This storage structure, which is built on columns, makes it possible to determine in parallel the ideal split for each attribute that is being taken into consideration. In addition, XGBoost uses a method that is based on data percentiles, in which only candidate splits sample is reviewed, and their gain is evaluated utilizing aggregated statistics, rather than examining all of the available candidate splits. This helps the algorithm produce more accurate results [9]. This resembles the subsampling of data at the node level that already exists in CART trees.

XGBoost based on thearea of Boostingtechniques of Ensemblelearning. Ensemble learning comprises of a group of predictors that are numerous models to deliver greater prediction accuracy. The earlier modelserrors in boosting technique are tried to be enhanced by succeeding models by applying certain weights to the models[10].

3.2. Firefly algorithm

The convolutional Firefly Algorithm; Yang created the Firefly method, a metaheuristic algorithm, in 2008 to solve optimization difficulties [18]. FA has proven to be better to other standard algorithms such as PSO and GA[19] by contrasting various functions of benchmark optimization. The three ideas that shaped FA's configuration are as follows [20]:

- The brightness of every firefly attracts another.
- The brighter the fireflies, the more attractive they are to other fireflies.
- The lower brightness levels of fireflies migrate to those with higher levels of brightness.

The three natural fireflies' behaviors prompted Yang to create the firefly algorithm, an optimization method. The behaviors of firefly and the creation of FA have a mutually beneficial relationship. In reality, the brightness of each firefly associated to each ideal solution will be determined by the optimal solution's fitness function. The search for and acquisition of many other fireflies generating greater brightness levels by fireflies having darker brightness levels is analogous to freshly created solution based on old solution with a superior fitness function. As a result, in FA, every old solution might be recreated multiple times depending on how bright it is in contrast to others. Consequence, depending on the comparing of fitness functions, just one new solution of every old solution is preserved.

Assume that each answer i (X_i) represents a firefly i location at the newest iteration. The spacing between the fireflies i and j is

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

calculated by utilize the following model whenever the fitness functions of solution i is greater than that of second solution j.

$$r_{ij} = \sqrt{(x_i - x_j)^2} \tag{3}$$

The modified distance is then used to calculate a newer attraction by substituting it into another (4). Then, corresponding to the creation of a new solution of the ^{i}th solution, thenewer location for the ^{i}th examined firefly may be calculated. The technique for making a newer solution is conducted in the following manner (5)

$$\beta = \beta_0 e^{-\gamma r_{i,j}^2} \tag{4}$$

$$X_{i,j_{new}} = X_i + \beta.rand.\Delta X_{i,j} + rand \tag{5}$$

where rand is the number of possible solutions at random. The attraction at distance of zero is i and β_0 and is generally set to 1. The following model determines $\Delta X_{i,j}$, which is an updated step size, and X_j is a solution with a lesser function of fitness than X_i.

$$\Delta X_{i,j} = (X_j - X_i) \tag{6}$$

For the ^{i}th solution, equations (3), (4), and (5) are calculated till the solution with a lower fitness function is not found. In conclusion, based on the fitness comparisons among the current population with solution i and other solutions, we can have one, more than one, or no new solution for each solution i. The following term can be used to describe the statement.

$$X_i^{new} = \begin{cases} X_i, & if\ X_i\ is\ X_{Gbest} \\ X_{iGbest}^{new}, & if\ X_j\ is\ X_{Gbest} \\ X_{ij}^{new}, & with \end{cases} \tag{7}$$

If the studied solution i is also the best global solution, no new generated solution for the equation of the first term(7). In the second scenario, just one new generated solution, X iGbest^new, if the evaluated answer is the second best and X_j is the population's overall best solution, XGbest. In other circumstances, it signifies that if X_i is the worst option, it is the third best or worse than the third greatest alternative, there will be two to (Npop – 1) new solutions X ij^new. In this case, the fitness function values comparisons will be used to evaluate the series of new solutions for solution i as well as the greatest one with the smallest fitness will be chosen, others are eliminated, but Gbest is kept.

4. Proposed model

Political regime classification is widely regarded as just a time consuming and difficult procedure. As a result, this study presented a new classifying solution for political articles. The proposed method sought to assess the efficacy of techniques of the established sentiment classification in relation to our novel compilation of political web posts. This research planned to use machine learning techniques that included support vector machines and evaluated their potential application in this field of study.

4.1 Dataset collection

The researchers used the corpus of 511 in the study. For this purpose, They gathered raw data from a variety of sources, including newspapers, websites blogs, and social media. As shown in Table 1, this dataset included three labels that covered topics such as politics.

Table 1: Describes this dataset

Post label	Number of posts

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

Reform Party	180
Conservative Party	158
Revolutionary Party	173

4.2. preprocessing

To prepare and clean the text for further classification, the data must be preprocessed. Online writings typically have lots additional noise and useless information like adverts, HTML tags, or scripts, which makes it difficult to extract words. Additionally, it is inaccurate to determine whether special Arabic characters, improper spellings, or punctuation marks are present. A significant number of Arabic words also have no impact on the text's orientation. By maintaining those words, it became more complex to be classified because each word was given its own dimension. There are several steps in the classifying process. In the first step, normalization, the words' signs are removed, and the identical word is converted into the symbol (◌ُ,◌ْ,◌َ,◌ِ,◌ّ,~,◌ً,◌ٍ,◌ٌ). The Arabic language is purified of all *harakat* in this process. After that, several Arabic characters, such as(——) is called the (tatwill) character, underwent changes. For example, (سياســـــة) became (سياسه). Additionally, some alphabets are changed, such as (إ, أ , ٱ , آ , ا) becoming (ا), (ى) becoming (ي); and (ئ,ؤ) becoming (ء) or (ة) became (ه). The researcher eliminated <3 alphabetic words throughout the normalization phase. By comparing them to the list of stop words in step two, the stopwords function aids in the filtration and deletion of undesirable Arabic words or tokens.The researchers have described 2 different lists, i.e., one built-in and in the second list, they added unwanted words as words which were unimportant for sentence formation. The stemming is the final step.

Stemming [21] was a vital stage in the pre-processing process utilized for Natural Language Processing (NLP) and sentiment analysis. The light stemmer strips all words of their suffixes and prefixes. Every word is reduced to its stem form through stemming, which also reduces the corpus data size to a smaller dimensional space. 4 major steps are used for Arabic stemming [22], out of which the Light stem was used in this study[23].

4.3. Feature extraction

The BOW model is the most widely used representation approaches in text classification [24], as well as the conversion of word docs to vectors [25]. The BOW approach is used to extract data from text sources. XGBoost algorithm can be trained using these characteristics. It develops a unique words vocabulary found in the training set's papers. In simple term, it is a set of words that are used to express a sentence with a word count and are not always in the same order. Words with higher weight values are given more weight, whereas words with lower weight values are given less.

To extract unique terms from the dataset in order to generate frequency dictionary. This means indicating the frequency of each word, which in this case will be determined by methods likeTerm Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF [26].

We must compute the TF-IDF and TF of every document word after getting vectors throughout all words. In (8) equation, $TF(w)$ refers to the document's word frequency w, $count(w)$ and $count(w_n)$ denotes the word w number in corpus and n refers to the number contained word w samples in corpus[27].

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

$$TF(w) = \frac{count(w)}{\sum count(w_n)} \quad (8)$$

In equation (9), $IDF(w)$ refers to the word w of the inverse file frequency, df refers to the samples number contained word w in corpus, and N refers to overall number of corpus texts [28].

$$IDF(w) = log\left(\frac{N}{df}\right) \quad (9)$$

This scheme aims to predict the product of TF-IDF. Formula (10) calculate the TF-IDE based on the equations (8) and (9).

$$TF - IDF(w) = \text{TF} * \text{IDF} \quad (10)$$

The BOW model advantages are; It's easy to comprehend and use, offers a lot of customization options for exact textual content, and, lastly, it's used to determine the words importance. BOW has also proved successful in forecasting challenges such as language models and classification [29]. Moreover, there are a few drawbacks to employing BOW. The first is semantic meaning [30] The basic BOW approach ignores the document's word meaning (Discard the order of the word: the same word could be used in many places nearby or on context words). Secondly, consider the size of the vector [31] could be a huge this results in a great deal of time complication.

Guenther and Sanderson [32] and Peng et al [33] proposed the N-gram as a text feature. A n-gram is a window with a length of n that slides over the text. As seen in the example in Figure, n-grams could work at the word, character, or even statement level, n-gram employed for sentiment analysis [34]. N-grams have three kinds which are: trigram, bigram, and unigram [35, 36]. Several features of natural language processing could be shown by n-gram [37, 38].

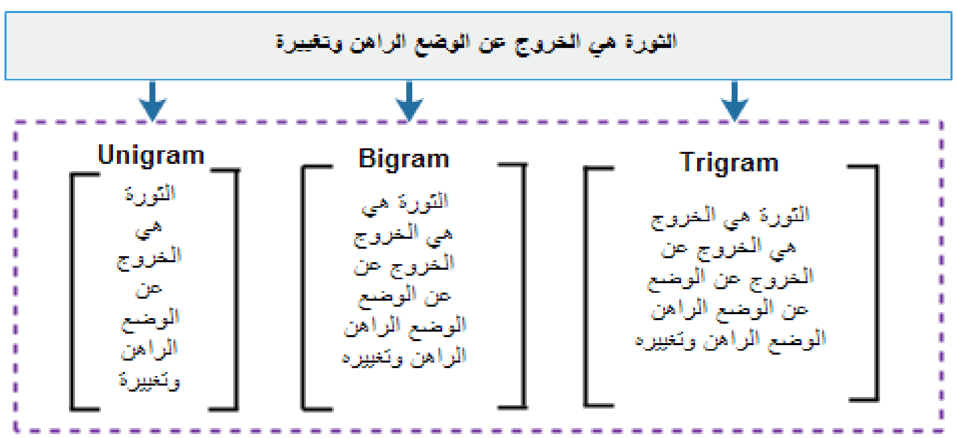

Figure 1: Example of Trigram, Bigram, and Unigram

The n-gram model has the following advantages: it is easy and simple to compute; it does not require encoding; it predicts words effectively when trained on huge amount of data; and it is also recommended to noisy text [39]. Although, the n-gram has two drawbacks: firstly, its interactions are not stored. Secondly, the lack of data at low frequency has an impact on the n-gram quality [40]. In this study used three grams which are unigram, bigram, and

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

trigram with two feature extraction TF and TF-IDF.

4.4. Feature selection

Feature selection is a keytopic inmachine-learning that has a substantial effect on the model performance. The data attributes utilized to train the machine-learning algorithm have a major impact on the final outcomes. Its purpose is to select a subset of the original collection of features which have the same significance but no loss in information. The existence of unnecessary data characteristics may degrade the proposed modelaccuracy and cause the model to train on irrelevant features. In numerous studies, the feature selection strategy is a crucial preprocessing step that leads to much more exact results. Feature selection has a number of advantages, including the following:

1. Deleting redundant, noisy and irrelevant data leads to lessening the run-time and reduce required storage media.
2. Improves Modelling Accuracy: Less confusing data means better modelling accuracy.
3. Training Time Shortening: With less data points, the complexity of the algorithm is reduced, and algorithms are training faster.

Xin-She Yang created the firefly algorithm (FA) while working in Cambridge in 2008, and it was published in Yang (2009). It created a method for optimizing functions with numerous optima by modifying the firefly swarm's behavior. To stimulate exploration of the solution space, it specifically made advantage of the idea that the individual fireflies brightness attracted them together and a randomization factor. Following the firefly algorithm publication, numerous studies analyzing and modifying the algorithm as well as accounts of its effective use in solving numerous real-world issues have been published.

The Firefly Algorithm (FA) is a metaheuristic algorithm that is used to optimize the machining settings. It was inspired by the flashing firefly's behavior and the bioluminescent communication phenomenon.

Firefly emit light or flash from the back of their body in order to keep predators away from them or to attract Firefly of the same kind to form groups in a specific direction. The goal is to search for food. Communicating with some of them is not through speech, but through the light emanating from each Firefly, as each Firefly follows the Firefly closest to it. The process of determining proximity is done by distance, when the distance is short, the light is stronger. In other words, Firefly are subordinate to the closest Firefly to them, which emits a stronger light than the others, and they are also followed by other Firefly, as they are the closest to them, and thus groups are formed.We proposed firefly optimization for select best words, Fitness function equation (11).

$$FT = \left(\frac{TP + TN}{TP + FN + TN + FP}\right) * 100 \qquad (11)$$

When FT is high then will be better, equation (11) used to calculate accuracy. The used

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

parameters in this research illustrated in Table 2.

Table 2: Firefly parameters

Parameter	Value	Description
Gamma	1	
Beta	0.20	
Alpha	0.25	
Pop	5, 10, 15, 20	
Iteration	25, 50, 75, 100	

Figure 2 present the flowchart for select best words from political Arabic post.

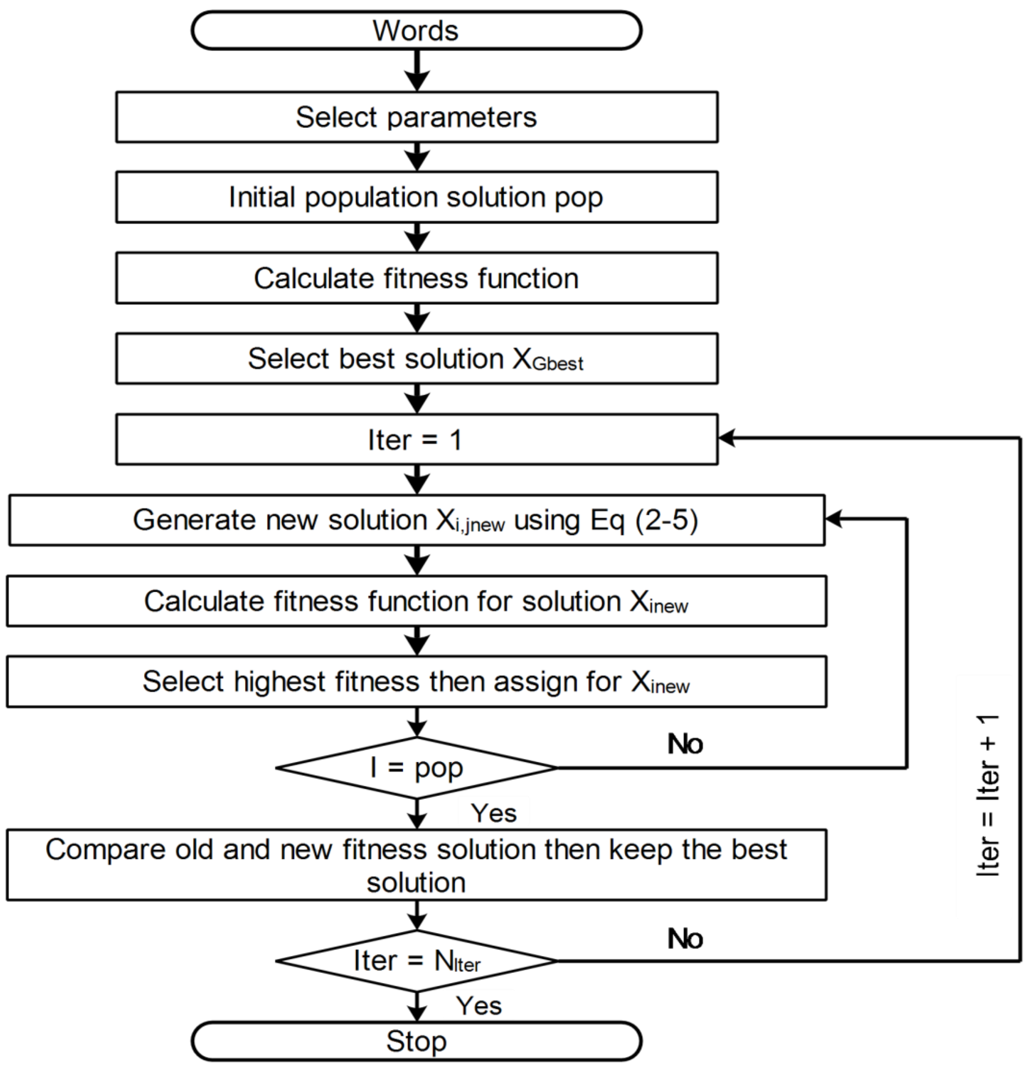

Figure 2: Firefly Flowchart for Select Best Words

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

The workflow of Firefly Show in Figure 1, Firefly read the words and select the parameters, after then, the process of initial population solution begins, when initial solution used, (FA) calculate the fitness function and selected the best solution $\mathbf{X_{Gbest}}$. Iteration begin from 1 to $\mathbf{N_{Iter}}$ to improve the objective function, generate new solution $\boldsymbol{X_{i,j_{new}}} where \boldsymbol{X_{i,j_{new}}} = \boldsymbol{X_i} + \boldsymbol{\beta}.\boldsymbol{rand}.\Delta \boldsymbol{X_{i,j}} + \boldsymbol{rand}$, calculate the fitness function for solution $\boldsymbol{X_{inew}}$ and select the highest fitness then assign for $\boldsymbol{X_{inew}}$, if the fitness is not equal to Pop, then return to generate new solution , but if the fitness is equal to Pop, algorithm will compare between then old and new solution and keep the best, finally, if the iteration ended, the algorithm will stop, else the iteration will increase by one and return process.

4.5. Evaluation Metric

In accordance with the confusion matrix, a number of measurements could be employed to assess how well the model performed in terms of accuracy[41]. These measurements determined by f-score, precision, recall, and accuracy, which are displayed in Table 3.

Table 3: Evaluation metric

Metric name	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{\mathrm{TP}+\mathrm{FN}}$
Precision	$\frac{TP}{TP+FP}$
F1-score	$2*\frac{Recall*Precesion}{Recall+Precesion}$

5. Results

5.1 Number of Features Extraction

Table *4*shows each classnumber for our dataset. Each class number of three grams is presented as shown in

Table *4*. It is crucial to determine in which gram the similarities and non-similarity number can maintain the same value in this section. Unique words that fall into a unique class are referred to be similar, whereas repeated or unique terms are referred to as non-similar. The dataset is split by similarity and non-similarity throughout the training phase. Since there are many posts in the training phase, it was chosen, and 70% of the entire dataset was used for training. Training is mostly used to mix words and analyze relationships between postings in a dataset with unique numbers.

This section displays the feature extraction number for XGBoost.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

Table 4: Feature Extraction Using three Grams

Class	**Numbers of each gram**		
	unigram	bigram	trigram
Conservative	2913	18631	21626
Reform	3152	26092	32521
Revolutionary	3616	26542	31614
Total unique words	5389	65392	85327

For each three grams, the total amount of unique words in the Revolutionary, Reform, and Conservative classes was taken into account. Due to the fact that they only have a one vector matrix for all classes, those words could support TF and TF-IDF, and as a result, they have been utilized in the algorithm of XGBoost. The high difference between classes was in unigram only shown in

Table *4*. Unigram and XGBoost learning can well working. It surpasses other grams in quality. Bigram is also seen to be better than trigram.

Due to the high number of the class type in the chosen posts, the revolutionary class in our experiment had the highest number of trainings among the three classes. The revolutionary class's three grams—the unigram, bigram, and trigram—acquired the most trainings. When bigram and trigram were used, the vector size grew correspondingly. Training time may be required for this XGBoost method. As demonstrated in

Table *4*Trigram brought the number between classes closer. Due to the same value of the vector, it is shown that trigram machine-learning almost produces the same accuracy for XGBoost.

5.2 XGBoost Algorithm

The main target of utilizing XGBoost algorithm was to compare their performance between different feature extraction and select best feature and to gain better results. In this regard, the specific number of posts was marked to each goal value in the chosen datasets. Due to the fact that the final selection dataset sample was limited to 511 posts divided into three classifications, the choice represents a trade-off. The primary goal of this experimental study was to achieve high accuracy and performance in political Arabic posts that will be extremely valuable for governmental issues like security and educational conditions. For each of the selected approaches, including the performance assessment metrics methodologies, this type of research introduced the problem of acquiring training sets and testing sets for multi-class label classification systems.

In this instance, the XGBoost algorithm was presented with the two feature extractions TF and TF-IDF. On our dataset, these methods were used. Results from this experiment were acquired by measuring the confusion matrix and visualizing them using techniques like ROC and precision-recall. The best feature frequency number and best feature extraction were then chosen by comparing the various

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

types of feature extraction numbers using the **Table** *5*and testing technique (TF or TF-IDF). **Table 6**show the TF and TF-IDF feature extraction with three grams.

Table 5: TF feature extraction with unigram, bigram, and trigram

n-gram	Class	Precision	Recall	f-score	Error number of posts	Accuracy
Unigram	Conservative	0.92	0.98	0.95	6	96.104
	Reform	0.96	0.94	0.95		
	Revolutionary	100	0.96	0.98		
Bigram	Conservative	0.93	0.79	0.85	25	83.766
	Reform	0.74	0.86	0.80		
	Revolutionary	0.87	0.86	0.86		
Trigram	Conservative	0.76	0.27	0.40	72	53.247
	Reform	0.73	0.38	0.50		
	Revolutionary	0.45	0.89	0.60		

Table 6: TF-IDF feature extraction with unigram, bigram, and trigram

n-gram	Class	Precision	Recall	f-score	Error number of posts	Accuracy
Unigram	Conservative	0.94	0.92	0.93	7	95.455
	Reform	0.94	0.96	0.95		
	Revolutionary	0.98	0.98	0.98		
Bigram	Conservative	0.93	0.79	0.85	27	82.468
	Reform	0.72	0.86	0.78		
	Revolutionary	0.87	0.82	0.84		
Trigram	Conservative	0.86	0.25	0.39	72	53.247
	Reform	0.73	0.38	0.50		
	Revolutionary	0.45	0.91	0.60		

The best accuracy was obtained from the XGBoost method as shown in**Table 7**. The best feature extraction that was used for XGBoost is TF as shown achieved 2 time than TF-IDF. XGBoost algorithm at unigram and bigram achieved good accuracy when using TF. That is why the vote was selected for them TF. The vote for TF was achieved all times for all number of features. The total points of TF achieved higher than that of TF-IDF.

Table 7: Select Best Feature with unigram, bigram, and trigram

Number of grams	Feature extraction		Vote
	TF	TF-IDF	

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

Unigram	1	0	TF
Bigram	1	0	TF
Trigram	-	-	TF& TF-IDF
Total	2	0	TF

The feature limitation and gram size are connected. The constraint is once the gram size is large, the vector size will also be large. A problem known as the time-consuming limitation is based on by large vector sizes the following section present how to select best words for building feature using firefly algorithm.

5.3. Firefly Algorithm

In this section, two types of features with firefly to select best feature are presented. The firefly was used with two vectors such as TF and TF-IDF. The present study improved feature selection by used firefly as shows in

Table *8*.

Table 8: Select best feature using firefly with unigram

Feature extraction	Number of features	Class	Precision	Recall	f-score	Error number	Accuracy
TF	210	Conservative	0.96	0.98	0.97	3	98.052
		Reform	0.98	0.98	0.98		
		Revolutionary	100	0.98	0.99		
TF-IDF	219	Conservative	0.90	0.98	0.94	6	96.104
		Reform	0.98	0.92	0.95		
		Revolutionary	100	0.98	0.99		

As shown in

Table *8*both feature extraction with unigram achieved good accuracy but with TF achieved higher than TF-IDF. Finally, the best feature extraction was TF where the number of words was only 210 that will be help to sole the limitation of feature extraction where select all words.

Table *9*Table 9 contrasts our suggested method with alternative approaches.

5.4. Proposed Model Comparison with Other Methods

In this section, we compared between the best XGBoost with firefly algorithm and other methods.

Table 9: Comparison with other methods

Reference	Objective	Dataset	Method	Accuracy	Limitation
[11]	Urdu news	Private	Using	75.44%	The text classification

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

	classification	dataset	XGBoost algorithm		that was developed for other languages that are similar to Urdu cannot be applicable to Urdu, because there is not enough elimination between words for it to be useful.
[13]	Automated Amharic News Categorization	CORD-19	Using XGBoost algorithm	84	Unable to reduce the dimensionality of a high dimensional feature space, it is hard to identify the number of filters that should be used for a certain task, as well as the suitable size of each filter.
[42]	Learning through active learning for sentiment analysis	Social Media	Using XGBoost algorithm	88.5	There is neither a big annotated dataset nor any tools or models that have been pre-trained for Telugu. To generate a word embedding model from Telugu data, essential preprocessing is required. It is difficult to locate a labeled training set because human annotation is time consuming and inefficient in terms of cost because it is cost-ineffective.
[43]	Unreliable Medical ArticlesDetectionon Websites of Thai	Collected from Thai Websites	Using XGBoost algorithm	90.60	Few numbers of the features
[44]	The Sentiment Analysis of Ukrainian text Model	gathered from Google Maps comments	Using XGBoost algorithm	90.76	Processing of a bigsize of unstructured information is an reallycomplex task, because the content of today's Internet is quite

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

	of Services Providers' Feedback				suitable for human perception, but remains difficult to access for machines.
[45]	Analysis of the Sentiment Analysis in Bengali Regarding E-commerce	obtained Bangladesh e-commerce website datasets from "Daraz,"	Using XGBoost algorithm	90.56%	Insufficient study has been done in the Bengali language to do sentiment analysis of customer reviews of products.
[46]	SentiXGboost: enhanced analysis of user comments on social media platforms using an ensemble of XGBoost classifiers	MSWASR	Using XGBoost algorithm	87%	Detecting similar questions is a research and industrial problem that has not been solved completely especially for the Arabic language
Proposed model	Sentiment analysis for Political Arabic post	Using political Arabic post dataset	Using XGBoost algorithm with firefly	98.052	-

5.5. Discussion

XGBoost was first utilized in this investigation with two feature extractions, TF and TF-IDF. In this study, political Arabic post datasets with

Table ***5*** and

Table ***6***.As indicated in section 5.2, the best feature extraction was TF.

Second, the proposed method was based on firefly algorithm to tackle the best number of features. In this study, one vector was used three classes were used. Three grams were used for each feature (TF or TF-IDF), which are unigram, bigram, and trigram. These features set with XGBoost algorithm was built as illustratedin

based on firefly selection. The firefly used unigram because it was the best according to section 5.2 as illustrated in**Table 7**. The comparison presented in section 5.3 shows that the firefly method was very good with TF feature extraction.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

The political discussion orientation on the three classes (Revolutionary, Conservative, and Reform) was observed in our experiment. A subset of this data was categorized for sentiment in order to study it as a whole as well as by classifying users based on their behavior in posting. The method was then developed post orientations identification. T hrough the study, as shown in Figure 3, the best accuracy was attained with the TF feature and typically with the first unigram, as indicated in Table 6. Throughout the practical portion of this work, the ROC and precision-call were utilized. The percentage of each class accuracy could be presented.

The proposed method's effectiveness could be seen when applying the dataset D2 and when making comparison with themethod of firefly. Previous work shown in **Figure 3**, achieved the three grams average accuracy, while the proposed method used firefly as feature selection with unigram with TF.

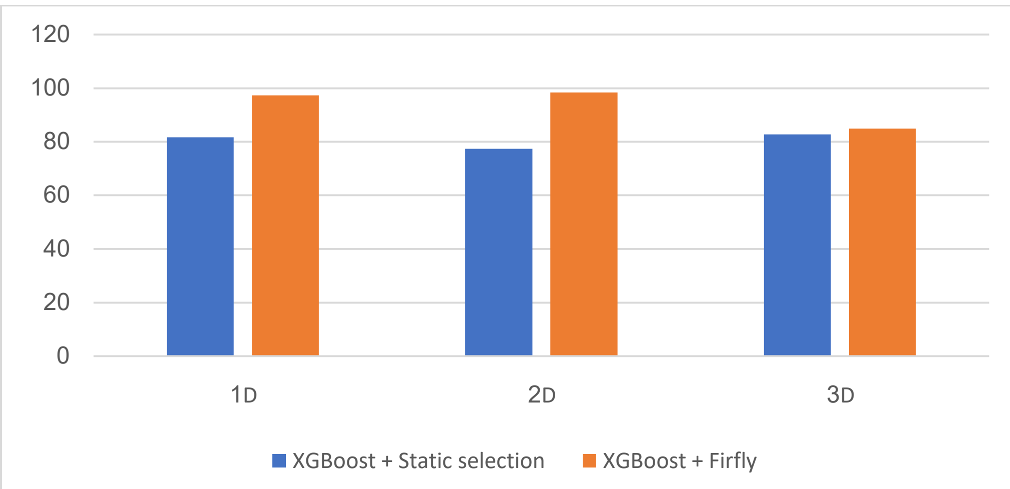

Figure 3: ThisResearch Method and Machine LearningComparison

Figure 3shows thatfor D1, D2, and D4, the firefly was much better from the static selection method. But, for dataset D3, the performance of statistic selection was so close to the firefly selection method with accuracy of 82 and 84 percent for the static selection and the firefly selection methods respectively. This is because the XGBoost method works well with root stemming.

Conclusion

The application of the Firefly algorithm to an original collection of datasets, which was derived from web posts, was the primary focus of our work. We provided evidence to support the hypothesis that an XGBoost classifier is sensitive to the class structure of the dataset. The purpose of this study was to evaluate the size of the dataset impact on the classification accuracy and efficiency of the XGBoost algorithm specifically by choosing the features that are most relevant to the problem. Make use of the TF along with the word count vector 210 for a very high level of precision (98.052 percent). In the future, we will be able to decrease the size of vectors by utilizing a variety of feature selection approaches such as PSO, the genetic algorithm, and others.

Reference

[1] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138-1152, 2011.

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1-167, 2012.

[3] C. Quan and F. Ren, "Unsupervised product feature extraction for feature-oriented opinion determination," *Information Sciences*, vol. 272, pp. 16-28, 2014.

[4] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Stream-based active learning for sentiment analysis in the financial domain," *Information sciences*, vol. 285, pp. 181-203, 2014.

[5] D. H. Abd, A. R. Abbas, and A. T. Sadiq, "Analyzing sentiment system to specify polarity by lexicon-based," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 283-289, 2021.

[6] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Political articles categorization based on different naïve bayes models," in *International Conference on Applied Computing to Support Industry: Innovation and Technology*, 2019, pp. 286-301: Springer.

[7] S. Alterkavı and H. Erbay, "Novel authorship verification model for social media accounts compromised by a human," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13575-13591, 2021.

[8] O. AlZoubi, S. K. Tawalbeh, and A.-S. Mohammad, "Affect detection from arabic tweets using ensemble and deep learning techniques," *Journal of King Saud University-Computer and Information Sciences*, 2020.

[9] A. N. Azhar and M. L. Khodra, "Fine-tuning pretrained multilingual bert model for indonesian aspect-based sentiment analysis," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, 2020, pp. 1-6: IEEE.

[10] M. S. Bhatti, A. Ullah, R. Latip, A. Sohail, A. Riaz, and R. Hassan, "Benchmarking Performance of Document Level Classification and Topic Modeling," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 71, no. 1, pp. 125-141, 2022.

[11] M. Bobur, K. Aibek, B. Abay, and F. Hajiyev, "Anomaly Detection Between Judicial Text-Based Documents," in *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, 2020, pp. 1-5: IEEE.

[12] D. Endalie and G. Haile, "Automated Amharic news categorization using deep learning models," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[13] D. Godavarthi, "Classification of covid related articles using machine learning," *Materials Today: Proceedings*, 2021.

[14] R. H. Hama Aziz and N. Dimililer, "SentiXGboost: enhanced sentiment

analysis in social media posts with ensemble XGBoost classifier," *Journal of the Chinese Institute of Engineers*, vol. 44, no. 6, pp. 562-572, 2021.

[15] M. I. Hossain, M. Rahman, T. Ahmed, and A. T. Islam, "Forecast the Rating of Online Products from Customer Text Review based on Machine Learning Algorithms," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 6-10: IEEE.

[16] H. Karayiğit, Ç. İ. Acı, and A. Akdağlı, "Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods," *Expert Systems with Applications*, vol. 174, p. 114802, 2021.

[17] Z. Li, Q. Zhang, Y. Wang, and S. Wang, "Social media rumor refuter feature analysis and crowd identification based on XGBoost and NLP," *Applied Sciences*, vol. 10, no. 14, p. 4711, 2020.

[18] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *International symposium on stochastic algorithms*, 2009, pp. 169-178: Springer.

[19] S. Fong, S. Deb, X.-S. Yang, and Y. Zhuang, "Towards enhancement of performance of K-means clustering using nature-inspired optimization algorithms," *The Scientific world journal*, vol. 2014, 2014.

[20] T. T. Nguyen, N. V. Quynh, and L. Van Dai, "Improved firefly algorithm: a novel method for optimal operation of thermal generating units," *Complexity*, vol. 2018, 2018.

[21] M. Mustafa, A. S. Eldeen, S. Bani-Ahmad, and A. O. Elfaki, "A Comparative Survey on Arabic Stemming: Approaches and Challenges," *Intelligent Information Management*, vol. 9, no. 02, p. 39, 2017.

[22] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 275-282.

[23] D. H. Abd, W. Khan, K. A. Thamer, and A. J. Hussain, "Arabic Light Stemmer Based on ISRI Stemmer," in *International Conference on Intelligent Computing*, 2021, pp. 32-45: Springer.

[24] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43-52, 2010.

[25] D. Sarkar, "Text Analytics with Python," 2016.

[26] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Political Arabic Articles Classification Based on Machine Learning and Hybrid Vector," in *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, 2020, pp. 1-7: IEEE.

[27] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Classifying political arabic articles using support vector machine with different feature extraction," in *International Conference on Applied Computing to Support Industry: Innovation and Technology*, 2019, pp. 79-94: Springer.

[28] L. Khreisat, "Arabic Text Classification Using N-Gram Frequency Statistics A

Comparative Study," *DMIN*, vol. 2006, pp. 78-82, 2006.

[29] L. Carstens and F. Toni, "Using argumentation to improve classification in natural language problems," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 3, pp. 1-23, 2017.

[30] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, "Subjectivity and sentiment analysis of Arabic: trends and challenges," in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, 2014, pp. 148-155: IEEE.

[31] S. ChandraKala and C. Sindhu, "Opinion mining and sentiment classification: A survey," *ICTACT journal on soft computing*, vol. 3, no. 1, pp. 420-425, 2012.

[32] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 482-491: Association for Computational Linguistics.

[33] F. Peng, D. Schuurmans, and S. Wang, "Augmenting naive bayes classifiers with statistical language models," *Information Retrieval*, vol. 7, no. 3-4, pp. 317-345, 2004.

[34] A. Dey, M. Jenamani, and J. J. Thakkar, "Senti-N-Gram: An n-gram lexicon for sentiment analysis," *Expert Systems with Applications*, vol. 103, pp. 92-105, 2018.

[35] A. Sharma, A. Nandan, and R. Ralhan, "An Investigation of Supervised Learning Methods for Authorship Attribution in Short Hinglish Texts using Char & Word N-grams," *arXiv preprint arXiv:1812.10281*, 2018.

[36] S. A. Taher, K. A. Akhter, and K. A. Hasan, "N-gram based sentiment mining for bangla text using support vector machine," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1-5: IEEE.

[37] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, 2012, pp. 90-94: Association for Computational Linguistics.

[38] D. Lin and X. Wu, "Phrase clustering for discriminative learning," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 2009, pp. 1030-1038: Association for Computational Linguistics.

[39] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*, 2002, pp. 1-8: Australian Computer Society, Inc.

[40] M. Flor, "A fast and flexible architecture for very large word n-gram datasets," *Natural Language Engineering*, vol. 19, no. 1, p. 61, 2013.

[41] J. K. Alwan, A. J. Hussain, D. H. Abd, A. T. Sadiq, M. Khalaf, and P. Liatsis, "Political Arabic articles orientation using rough set theory with sentiment lexicon," *IEEE Access*, vol. 9, pp. 24475-24484, 2021.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4079-4101
https://publishoa.com
ISSN: 1309-3452

[42] M. Parimala, R. Swarna Priya, M. Praveen Kumar Reddy, C. Lal Chowdhary, R. Kumar Poluru, and S. Khan, "Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach," *Software: Practice and Experience,* vol. 51, no. 3, pp. 550-570, 2021.

[43] C. Saengkhunthod, P. Kerdnoonwong, and K. Atchariyachanvanich, "Detection of Unreliable Medical Articles on Thai Websites," in *2021 13th International Conference on Knowledge and Smart Technology (KST)*, 2021, pp. 102-107: IEEE.

[44] K. Shakhovska, N. Shakhovska, and P. Veselý, "The sentiment analysis model of services Providers' feedback," *Electronics,* vol. 9, no. 11, p. 1922, 2020.

[45] M. T. Akter, M. Begum, and R. Mustafa, "Bengali sentiment analysis of E-commerce product reviews using K-nearest neighbors," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 40-44: IEEE.

[46] Y. Alotaibi *et al.*, "Suggestion mining from opinionated text of big social media data," *Computers, Materials & Continua,* vol. 68, no. 3, pp. 3323-3338, 2021.