

## A large-scale sentiment analysis using political tweets

Yin Min Tun, Myo Khaing

Faculty of Computer Science, University of Computer Studies, Mandalay, Myanmar

---

### Article Info

#### Article history:

Received Aug 19, 2022

Revised Apr 26, 2023

Accepted Jun 26, 2023

---

#### Keywords:

Apache flume

Big data analytic

Machine learning

Sentiment analysis Apache

Social media data

Spark

---

### ABSTRACT

Twitter has become a key element of political discourse in candidates' campaigns. The political polarization on Twitter is vital to politicians as it is a popular public medium to analyze and predict public opinion concerning political events. The analysis of the sentiment of political tweet contents mainly depends on the quality of sentiment lexicons. Therefore, it is crucial to create sentiment lexicons of the highest quality. In the proposed system, the domain-specific of the political lexicon is constructed by using the supervised approach to extract extreme political opinions words, and features in tweets. Political multi-class sentiment analysis (PMSA) system on the big data platform is developed to predict the inclination of tweets to infer the results of the elections by conducting the analysis on different political datasets: including the Trump election dataset and the BBC News politics. The comparative analysis is the experimental results which are better political text classification by using the three different models (multinomial naïve Bayes (MNB), decision tree (DT), linear support vector classification (SVC)). In the comparison of three different models, linear SVC has the better performance than the other two techniques. The analytical evaluation results show that the proposed system can be performed with 98% accuracy in linear SVC.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

### Corresponding Author:

Yin Min Tun

Faculty of Computer Science, University of Computer Studies

Mandalay, Myanmar

Email: yinmintun@ucsm.edu.mm

---

## 1. INTRODUCTION

Social media have become more essential and Twitter plays a vital role in campaigning during election time. Twitter is one of the most common and popular social media that give the freedom for people to share their opinions, thoughts, and beliefs in the world. Twitter is increasingly used by politicians, journalists, political strategists, and citizens as a large part of the network for the discussion of public issues. Governments and politicians always detect the social media network and amendments, and how people are responding to different policies, and acts. Some political scientists working with Google, Facebook, or precise large datasets may have to know about big data architecture and new distributed methods with the huge data sets. Political scientists can focus more on new software for data cleaning, data management, reproducible science, data lifecycle management, and data visualization. In the era of big data, data is collected from various sources, such as mobile devices and web browsers, and stored in various data formats. It cannot handle the traditional storage and analytics platform from various structured and unstructured data. Hadoop, a good platform for big data analytics, offers scalability, cost-efficiency, parallel processing, availability, flexibility, and fast and secure authentication. An open-source framework Hadoop comprises a storage part called Hadoop distributed file system (HDFS) and a processing part called MapReduce. Sentiment analysis (SA), one of the big data applications focuses on analyzing big data in various ways, and

includes identifying patterns and relationships, making informed predictions, providing actionable insights, and deriving insights. SA uses text analytics in analyzing steps with high velocity and a large amount of tweet data. Sentiment classification techniques can be roughly divided into machine learning approach, lexicon-based approach, and hybrid approach. The lexicon-based approach is built in a sentiment lexicon that has known the collection of data and precompiled sentiment terms. In this approach, the emotional dictionaries can classify the analysis of specific words and sentences from the tweet. The emotional vocabulary element of the dictionary is searched in the text, the emotional weight is calculated, and the aggregate weight function is applied. This technique is governed by the use of a dictionary consisting of previously labeled vocabulary. The text classification depends on the total score obtained from the emotional elements. The machine learning approach is used the sentiment analysis tasks in this system which collect data from the tweet to label text classification. Machine learning methods can be mainly applied to the emotional analysis by using supervised classification. It applies the famous machine learning (ML) algorithms to analysis the linguistic features. Sentiment lexicon is mainly organized by using both approaches which are the majority of methods called hybrid approach.

Political sentiment lexicon, a resource of the lexical intended for sentiment analysis, is a lexical element extreme word database at the presidential election time with their sentiment polarity for a political domain. Political lexicon has been widely used for analyzing the extreme opinion word in a web text. SentiStrength has deployed references for the extreme opinion word weight resource to categorize the orientation of the political domain lexicon. The building of the political sentiment lexicon is one of the challenging components in the approach of lexicon-based sentiment classification. There is no result to point out the extreme political words from the sentiment lexicon. So, the approach of lexicon-based political sentiment classification applies the political domain lexicon to classify the sentiment orientation of the sentence. The construction political sentiment lexicon can have an effect on the performance of the extreme political opinion words analysis and sentiment classifier. The performance of the political domain-specific lexicon is the best enhancement to classify the sentiment lexicon words from the political tweet content.

In this paper, political multi-class sentiment analysis (PMSA) is implemented on a big data analytics platform. In PMSA, political sentiment lexicon is constructed by extracting valuable information from a large source of data. Multi-class classification is performed with three different machine learning techniques. The proposed PMSA developed with four modules: data collection, data preprocessing, lexicon generation, and data classification.

## 2. RELATED WORKS

This section described some papers that are related to the proposed PMSA. Three parts of this section are political lexicon construction, political sentiment analysis and the works of the proposed system. The first part describes the related paper on political sentiment lexicon-generating approaches. In second, political sentiment analysis papers are described and the last parts summarized the proposed work.

### 2.1. Political lexicon construction

Vu and Thanh [1] proposed a lexicon-based SA system on social media networks. The authors implemented the lexicon-based approach which certain the sentiment dictionaries by utilizing the heuristic method and the data preprocessing. Their proposed method was effective because this method is a combination of popular lexicon-based sentiment analysis methods, the Liu method, and SentiWordNet. Moreover, the data preprocessing steps of this system filter the opinion-oriented word from the text data before the sentiment analysis conducts. The performance of this proved method was achieved better than any previous method with the lexicon-base method.

Almatarneh and Gamallo [2] presented a lexicon-based approach for searching extreme opinions. They used the unsupervised approach for searching extreme opinions which were based on the automatic new lexicon construction with the most positive and most negative words. The main purpose of this system is to assign a value to extreme opinions. They described automatically a method to create the lexicon of the extremely positive and negative words from labeled corpora. Their automatically created lexicons had been compared with the other previous lexicons by performing the account of some partitions.

Feng *et al.* [3] implemented an automatic sentiment lexicon generation approach for the reviews of mobile shopping. In this system, the authors proposed the automatic constructing approach for the sentiment lexicon of a specified domain by considering the relationship between product features and sentiment words in the reviews of mobile shopping. There are two main parts; the first part is selecting product features and sentiment words from the original reviews and then using categories to perform the dimension of sentiment. Second, sentiment words that are related to mobile shopping are classified or clustered into specified

categories to form a dimension of sentiment. Their generated lexicon is created by constructing the classification task of sentiment with the various products that are written reviews in both English and Chinese.

## 2.2. Political sentiment analysis

Political Sentiment analysis is one of the interesting SA systems through online political tweets for the prediction results of the election. Rohrschneider *et al.* [4] presented the political SA system during the election of the German Federal in 2009. The purpose is to specify a platform for sentiment analysis on Twitter data and also predicted the outcome results of the election. The authors determined by the classification of which a political party or politician is identified. They applied the LIWC2007 tool for sentiment extraction from related political tweets. LIWC was an accurate software of text analysis that was developed to reveal thoughts of people, cognition, personality, and emotions by representing text samples. Fujiwara *et al.* [5] concluded their system that the number of tweets was directly proportional to the winning chances of election.

Ringsquandl and Petkovic [6] developed SA on the campaign topics of presidential candidates in the Republican Party, USA. They presented the frequencies amalgamation of noun phrases and the pointwise mutual information (PMI) measure of their system with a limitation on aspect extraction. The authors described the semantic relationship between their topic holds and politicians. According to their experimental result, the accuracy of the aspect extraction in their system is improved. Elghazaly *et al.* [7] implemented the SA in the Egypt presidential election based on the classification of Arabic text using the WEKA application. They expressed their results that the highest accuracy result was achieved by using the naïve Bayes (NB) classifier with the lowest rate of error.

Caetano *et al.* [8] implemented the political sentiment analysis to identify the political user's classes and user's homophily during the American presidential election, in 2016. They collected the tweets data of 4.9 million from the 18,450 users and their network from August to November 2016. The author specified six types of user classes which are representing their sentiment words for Hillary Clinton and Donald Trump: whatever, Hillary supporters, Trump supporters, neutral, negative, and positive. Their experimental results it is a better homophily levels that supports the multiple connections, the similar speeches, or the reciprocal connections.

Ullab *et al.* [9] proposed the political sentiment analysis system for optimal searching from the presidential elections in the USA 2016 to prove that their features were more suitable in the election results prediction. Their applied features such as uni-gram, bi-gram, tri-gram and opinion words features were analyzed and compared by using the popular data mining approaches such as random forest (RF), artificial neural networks (ANN), and naïve Bayes (NB) classification methods. They implemented many preprocessing methods on the dataset to expose the well\_shaped dataset. Finally, they proved that their system found the unigram showing with a higher accuracy of 81%.

After studying the related research paper, the proposed system is considered a political multi-class sentiment analysis system for Twitter data. Political multi-class sentiment analysis constructs the political lexicon. The accuracy of the lexicon is evaluated by different political data. The implementation of PMSA is performed on the big data platform.

## 2.3. Proposed work

Enhancing the performance of sentiment analysis for the political domain is the major objective of this work. Twitter data analysis is related to the mining of text because most of Twitter posts are text messages. It encompasses the methodologies such as machine learning, natural language processing, and data mining to appropriately characterize measure, model, and mine meaningful patterns from political tweet large-scale data. This system's studies review the various methods for generating political resources during the presidential election season. To extract opinion words from political tweet content for sentiment analysis, the proposed system investigated the extreme opinion word of lexicon generation for political domain. This system considers opinion data in sentiment analysis. Extreme opinion word needs linguistic resources such as lexicons, corpora, and dictionaries to implement sentiment analysis in political tweets. Political tweet resources help to evaluate extreme opinion word sentiment analysis for election time. The proposed system is based on the creation of a political lexicon, classification of political opinion words, and political tweet content mining process. The system expresses the background theories on a big data analytic platform for developing political tweet sentiment analysis that is applied the initial experiments of this research and the proposed combined political lexicon generation and machine learning based for political tweet informal short text.

## 3. PROPOSED METHOD FOR CONSTRUCTING SENTIMENT LEXICON

Sentiment analysis is still having many challenges and researchers are trying to solve the problems from the various disciplines. The applications of sentiment analysis are being promised in various political and business industries. Therefore, researchers became to introduce a lexicon-based sentiment analysis approach for the election twitter data.

### 3.1. How to build political lexicon for political domain

Sentiment classification aims to automatically classify political retweet text into positive or negative sentiment. Machine learning-based, lexicon-based, and hybrid approaches can be used to classify sentiment. An important tool for determining the sentiment polarity of tweet user opinion is a sentiment lexicon. Methodologies, knowledge-based and corpus-based are commonly used to create sentiment lexicons. Due to the political system, there are domestic politicians in a large number of supporters and voters in their campaign places. They get political information and opinion that they want to elect and they inform their message on the internet. Sentiment analysis and opinion mining have been focused on many aspects related to opinion, namely polarity classification by making use of positive, negative, or neutral values. However, most studies have overlooked the identification of extreme opinions in spite of their vast significance in many applications. The political lexicon generation uses a supervised machine learning approach to search for extreme opinions, which is based on the automatic construction of a political lexicon containing positive, strong-positive, negative, strong-negative, and neutral words.

Political sentiment lexicon is a resource for sentiment analysis which is a lexical element extreme word database at the presidential election time off with their sentiment polarity for a political domain. Political lexicon can be widely used for analyzing the extreme opinion words in a web text. The political lexicon recognizes political domain-specific sentiment based on the opinion of the Twitter user. Sentiment scores (positive, strong-positive, negative, strong-negative, and neutral) are generated for the opinion words which are based on the political tweet terms. The total word occurrence is calculated by the term frequency-inverse document frequency (TF-IDF) method [10]. As the first step of this process, “parts of speech” for each word are extracted, and then ranks for each word are defined as part-of-speech (PoS) tags words. In the next step, these tags are normalized. In the final step, the total score for each multiplication vector word is computed with the initial score. Sentiment score calculation is used to calculate the score of sentiment on the preprocessed data to define the class label on data. In this approach, feelings, weight calculations, and functions of total weight are applied for emotional vocabulary dictionaries list searching in the text. This proposed technique is improved with the help of a pre-tagged lexicons dictionary and the TF-IDF method for the creation of an input tweet vector. In this system, the political Twitter stream data contains opinion words that help to determine sentiment about the political parties. Political lexicon is used by political parties to extract with the opinion words in the president’s election.

The political lexicon is the mainly effective way to express their feelings and opinions when they vote. The generated political lexicon applies on the political domain to classify the sentiment orientation of the sentence. In other terms, the construction political sentiment lexicon can have an effect on the performance of the extreme political opinion words analysis and sentiment classifier. The performance of the political domain-specific lexicon classifier is enhanced as the sentiment lexicon which includes the political words containing well-built sentiment to classify the opinion words the political tweet content. Extraction the extreme opinion words for political tweet content by using a dictionary-based approach intends to sentiment classification. The opinion seed word is a word utilized for collecting antonyms and synonyms as of the dictionary and applied for sentiment analysis. Furthermore, if the opinion seed words contain incorrect sentiment polarity, the sentiment classifier incorrectly classify the sentiments of political words. Automatic lexicon is generated by using the approach of constructing a lexicon from the trained data. The lexicon-based political sentiment lexicon generation is proposed for the political domain.

### 3.2. Knowledge-based

Knowledge-based is a graphic resource such as WordNet. The political lexicon is developed as a knowledge-based dictionary to find synonyms and antonyms of words. It finds closer not only the synonym words but also the antonym words on tweets, the less iteration is required to define a synonym between the words. Both studies use the relationship between words in a knowledge-based. The main idea of these methods is to manually collect the initial set of sentiment words and their orientations, and then it uses the knowledge-based to expand this by searching their synonyms and antonyms.

### 3.3. Corpus-based

These methods are based on syntactic reasons or patterns that occur in the first list of word senses to find other meanings in the larger corpus. Based on the corpus, this was found in the change of emotional polarity in the text. It gears the emotional polarity of a word in the corpus-based system to towards for the emotional polarity of its neighboring words. Both works are based on a corpus rather than a knowledge-based. The great advantage of corpus-based methods is the domain-specific words and their orientations found in the finding process.

### 3.4. Machine learning

Machine learning methods are based on annotated data classification into intended categories. These can be grouped into three broad categories: supervised, semi-supervised and unsupervised. When people do not annotate their data, individuals choose a vocabulary-based approach. Machine learning approaches are fully automated, convenient, and capable of handling large data collections. These methods require a training dataset to support classifier automation. It is used to develop a classification model for classifying feature vectors and a test data set for predicting class labels for unseen feature vectors. Most operations of the sentiment analysis use machine learning. In this system, three machine-learning methods were applied (multinomial naïve Bayes (MNB), decision tree (C4.5), and linear SVC), to classify the sentiments expressed in politicians' retweet sentences written in political tweet content.

The combined lexicon-based and machine-learning methods consider extracting extreme opinion word and constructing a classification model to evaluate the lexicon performance. Proposed system is available a unified framework in which lexical background information, unlabeled data, and labeled training sentences can be effectively combined. This system analyzes the political retweet content and BBC News political discourse sentiment sentences.

## 4. PROPOSED METHOD FOR POLITICAL SENTIMENT ANALYSIS

Alaoui *et al.* [11] presented a three-stage-based approach, which is the dynamic dictionary construction with the words' polarity on the given topic based on the selected hashtags. In their proposed research, related tweets with the 2016 US election are specified to the negative and positive classes. According to their evaluation results on traditional data analytics, high accuracy of prediction performance was achieved but they cannot provide improved accuracy on the big data analytics platform. In 2019, a sentiment analysis system based on a supervised machine learning approach was created for the general election of India with the political sentiment data on Twitter [12].

They used long short-term memory (LSTM) for the classification model and analyzed the experimental results with the other machine learning approach models. They do not consider big data analysis and do not perform multi-class classification. Hasan *et al.* [13] implemented the hybrid technique on the political analysis with two machine learning techniques (SVM and naïve Bayes).

In this research, they evaluated performance analysis on the two machine learning approaches, but they did not consider multi-class classification. Baltas *et al.* [14] presented a sentiment analysis tool for the microblogging messages analysis with their specified sentiment. In their system, naïve Bayes, decision trees, and logistic classification methods were used by MLlib for the classification outputs. According to their experimental results, the accuracy rate of the Naïve Bayes classification method is more achieved superior standard results than the other two methods but they also did not perform on multi-class sentiment analysis [15].

The author implemented sentiment analysis for personalized tweets recommendation and tweets classification using a specific domain seed list to classify the tweets. They got 96% of accuracy on this dataset when they implemented one million datasets on tweet [16]. Thelwall *et al.* [17] detected sentiment strength in the short informal text. In this system, the author optimized the sentiment strengths using the machine learning approach, and then they approved that their proposed system is better than the other baseline approaches.

Bouazizi and Ohtsuki [18] proposed multi-class sentiment analysis on the Twitter data for the challenges and performance of classification. They proposed a new model to present the sentiments and applied this model to describe the relationships of different sentiments and to discuss about the difficult task on the multi-class sentiment analysis. The available raw data are collected from various social media sources and the data types are unstructured, semi-structured, and structured. Among the various popular social media networks, the feature of Twitter is a combination of social networks and web blogs [19]. The rapid improvement of Twitter helps journalists, citizens, politicians, and political strategists. This fact supports for an important part of the media network with publicly mediated and political affairs [20].

### 4.1. The framework of proposed political sentiment analysis system

This framework has been applied for the implementation of the parallel and distributed execution system with the processing of batch data on the different data nodes. For this approach, Apache Hadoop has been utilized as the best framework for implementation [21]. On the other site, Apache Spark provides the achievement with more attractiveness and high performance by expanding Hadoop's abilities and permitting the processing of real-time stream data [22]. In this system, extracting the required information from political big data on social media is developed. This system is implemented the platform of big data analytics (Apache Spark) with a highest amount of velocity and a large amount of tweet data. To get high accuracy, this system

is developed using the hybrid technique of lexicon-based classification and machine learning techniques with classification.

There are four main modules in the proposed system: a collection of data, preprocessing of data, generation of political lexicon, and data classification module. The development of these modules is implemented on four layers: a data ingesting layer, a layer for storage, a layer for processing, and a layer for data analytics. Real-time stream data from Twitter is collected by using Apache flume and then this collection of ingested data is developed at Hadoop distributed file system (HDFS) using a memory channel. HDFS is located at the storage layer of this system and Spark is used at the processing layer for batch processing. The implementation of other modules is performed at the analytics layer. The hybrid approach (a combination of three machine learning approaches and lexical-based classification) is a very useful approach for sentiment analysis with optimal performance.

#### **4.2. Implementation of big data platform**

Political sentiment analysis of multi-class systems is developed using the platform of big data analytics with Spark streaming, HDFS, Spark MLlib, and Apache Flume [23]–[25]. The functions of the four layers are discussed in this section. The layer of data ingestion-At this layer, the collected stream data from Twitter is performed by Apache Flume. These are pushed to the HDFS-sink for offline processes. The receiver is set up by spark streaming as Avro-agent and the collected data is pushed into Avro-sink for the online processes.

The layer of data storage: To store reliable and usable collected data, HDFS is used in this storage layer. HDFS helps the architecture of master/and single name of node implements as the primary server. The operations of file namespace are performed by single name node as opening, renaming, and closing. The directories and files that define the block map to data nodes. In HDFS, stored data is implemented by data nodes.

The layer of data processing: In this layer, Spark and YARN Cluster Manager are used for large amounts of data parallel processing based on hardware clusters for reliable nature and fault-tolerant. In YARN, the YARN Resource Manager tracks the resources for node Manager. This node manager controls the resources of the slave node. At the same time, each slave node has one or more resources and each executor has a process of work. According to the schedule, a separate Java virtual machine (JVM) performs the task under the operator. The loaded data is separated as multiple handlers in resilient distributed dataset (RDD) and these partitions use the conversion. This Manager have to run simultaneously the multi-threaded and multiple tasks.

The layer of data analytics: in this layer, there are two main components: classification for online and training for offline. The processing of data, labeling for class, and model generation are implemented for offline training to create the model of classification. The upcoming real-time of tweets preprocesses are classified with the online classification. The analytic layer execution process is implemented with a distributed Spark engine. The sentiment classification of multi-class based on the domain-specific sentiment lexicon of the political lexicon was developed by the help of Spark.

As shown in Figure 1, political data is the associated activities to make the decisions for the groups as the distributed resources. To develop the useful extraction of efficient political information, this proposed system implementation has four steps: a collection of data, preprocessing for data, class labeling, and data classification for the sentiment. In the data ingesting layer, real-time stream data from Twitter is collected using Apache flume and then this collection of ingested data is developed at Hadoop distributed file system (HDFS) using a memory channel. HDFS is located at the storage layer of this system and Spark is used for the processing layer in the batch processing. The implementation of other modules is performed at the analytics layer. The hybrid approach (a combination of machine learning approaches and lexical-based classification) is very useful for sentiment analysis with optimal performance.

#### **4.3. Data collection**

In this presented system, stream data through Twitter are collected by Apache Flume with the configuration of batch streaming, and then these collected data are assigned to HDFS-sinks [21]. English tweets data are collected and filtered by using keywords. Twitter data and Apache Flume play the main role in data collection and briefing of Apache Flume descriptions.

#### **4.4. Data pre-processing**

As the ingestion of data includes the useless data and data duplication, these data need to preprocess and clean for an effective analysis system. The purpose of data preprocessing is the removing of tweet text feature selection, duplicated and noise data, stop words, and repeated characters. In this step, the classification job is simplified and the cost of processing is decreased for the training stage.

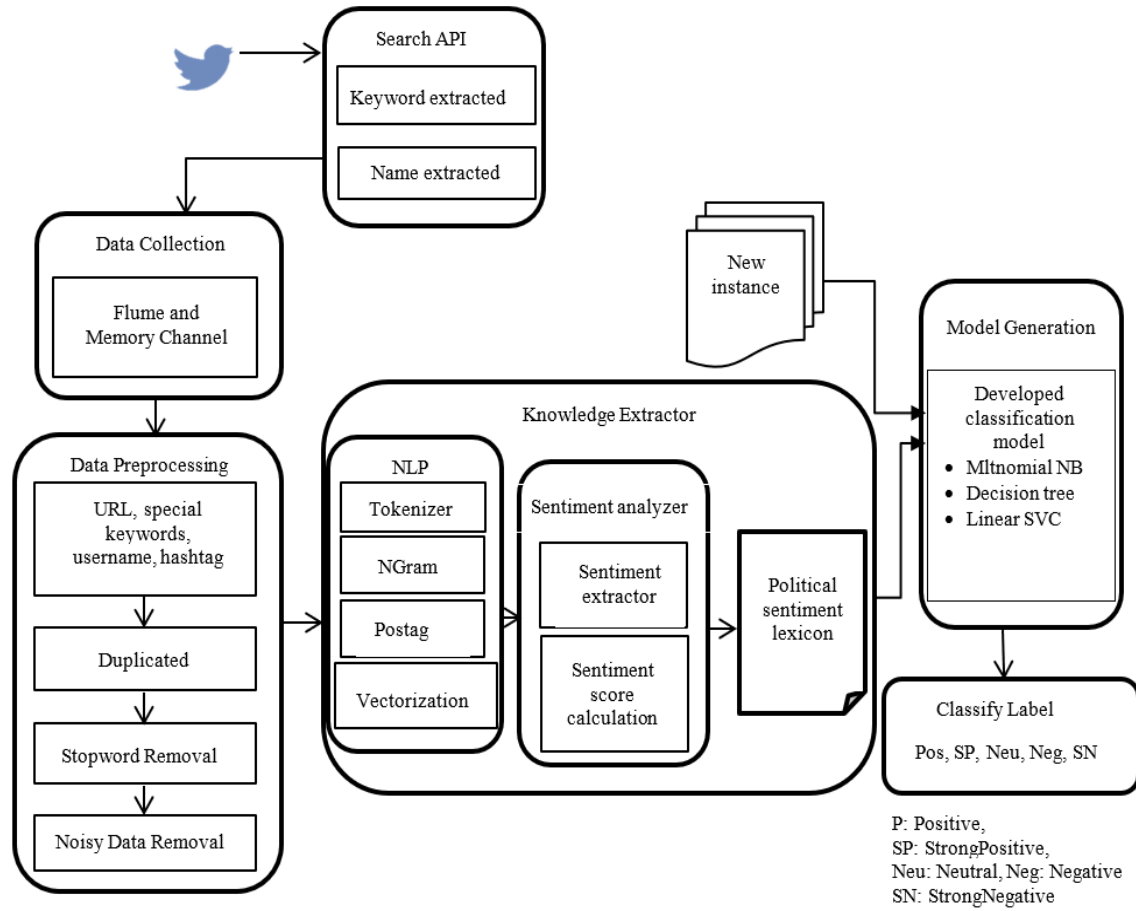


Figure 1. Political sentiment analysis system

**4.4.1. Tweet text feature selection**

The selected tweet features are used for sentiment analysis which describes the feeling and opinions of the Twitter user. One tweet stream record has various tweet attributes. Among them, the “text” attribute describes the opinion and feeling of the Twitter user. For the proposed system development, the values of the “text” attribute are selected because this system is the analysis for mood tracking of the people. English text attributes are used in this system. In the selection of values “text” attribute, the values of the “lang” attributes are checked and extracted whether the text language is English or not. If the “lang” attribute is Language in English, these are also selected for this system’s analysis.

**4.4.2. Removal of noisy data**

In tweet data, the useless information is noisy data for the tweet data classification. These data include hashtags and Website URL links where hashtags are replaced with the same word without hashtags by using adding character repetitions, white space, and @username. For example, #fun is replaced with fun. In this system, a non-alphabet replacement using space is developed.

**4.4.3. Removing of character repetition**

To remove duplicate tweets, the extracted tweet checking is used to determine which already exists and which does not. If there is data duplication in tweets, there is no use extra words, and these tweets are replaced with the duplicated tweets deletion by using the substring function. It is combined with an existing list of tweet data and analyzed when the feature extractions from tweets are already completed.

**4.4.4. Removing of stopwords**

Removing stopwords is an important technique for noise elimination in the classification of tweet text. Tweet data need to remove stopwords due to the inclusion of no meaningful information. The stopwords of this system not only depended on domain classification but also took manual examination of data into consideration.

#### 4.4.5. Negation handling

Negation handling is to identify the scope of negative and upturn the polarity of opinionated words that are effective by a negative. It supports significantly the accuracy of the classification. For example, the word “not good” in a phrase may represent negative sentiment that is follow by not into “not\_” + word.

#### 4.5. Political lexicon generation

In lexicon generation, the cleaning data is allotting into chunks with unigram, bigram, trigram, and n-gram words. And then it uses a JSON parser to split the parts of the POS tag. Extract the political opinion words (Adj, Adv, NN, and Verb) that are supported and related to the candidate. The TF-IDF method is used for counting. This method includes the weight to terms that recur frequently in a document as important words. Using TF-IDF, relevant words or keywords in a news text are found and then it gives the weight of word based on how many the recurrent features are in the political news.

#### 4.6. Sentiment score calculation

Each sentence is assigned by its appropriate polarity in accordance with the sentiment analysis criteria at the sentence level. This is typically accomplished by determining the polarity of individual words, and phrases, then combines them to determine the overall polarity of the political tweet sentence. The mood expressed in each political word was estimated by adding the polarities of the words and phrases in the political tweet content and considering each new piece as a sentence.

Political word scores on the content of political news tweets were calculated by using the sentiment extraction operator. This operator gives the final result in terms of sentiment. Texts with sentiment scores between -1 and -5 are considered negative and strong negative, while texts with sentiment scores between +1 and +5 are considered positive and strong positive. This operator provided the accurate results by using the political lexicon. We also use the sentiment score function which is based on (1)-(5) to calculate the total sentiment score for political news content on tweets.

$$P(S) = P(W) \quad (1)$$

$$P(S) = P(B) * P(W) \quad (2)$$

$$P(S) = (-1) * P(W) \quad (3)$$

$$P(S) = \sum_{i=1}^n P(W)_i \quad (4)$$

$$P(Avg) = \frac{P(S)}{n} \quad (5)$$

where  $P(S)$  is total polarity of a sentence,  $P(W)$  is polarity of word,  $P(B)$  is polarity of booster word,  $P(Avg)$  is average polarity of the sentence and  $n$  is number of sentiment words.

#### 4.7. Class labeling

Lexicon-based political generation is implemented in the training data classification to define the label. The procedure of class labeling process is described in the following. The class label of a sentence depends on the total polarity of a sentence.

##### Procedure: Class labeling

Input: raw tweets

Output: tweets, class label

1. Begin
2. Calculate total polarity strength on each sentence by applying generated political lexicon with SentiStrength\_Data
3. If (score>1) then print "StrongPositive"
4.     Else if (score>0 && score<=1) then print "Positive"
5.     Else if (score==0) then print "Neutral"
6.     Else if (score>=-1 && score<0) then print "Negative"
7. Else if (score<-1) then print "StrongNegative"
8. emit (tweets, class label)
9. End

#### 4.8. Classification model

In the final step, the development of classification models is used for class labels and feature vectors. The optimal model with the best accuracy was selected from the other developed classification



models of this system [26]. In the training data and testing data of this proposed system, the three different classification methods (MNB, decision tree, and linear SVC) are implemented [27].

#### 4.8.1. Multinomial naïve Bayes (MNB)

Bayes theorem-based multinomial naïve Bayes algorithm is a probabilistic learning method, which is mostly applied in the research area of NLP. MNB evaluates the probability of each tag for a given sample and then returns the tag with the highest probability as the output. Naïve Bayes classifier is a collection of many algorithms in which all the algorithms share one common principle, and that is, each classified feature is not related to any other feature.

- Maximum likelihood

$$\theta_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad (6)$$

- Posterior probability

$$P(y|x_i) = P(y) \prod P(x_i|y) \quad (7)$$

where  $\theta_{yi}$  is the probability  $P(x_i|y)$  of feature  $i$  appearing in a sample belonging to class  $y$ .

#### 4.8.2. Decision tree (DT)

Decision tree-based classification model, also known as a statistical classifier, is an approach for the classification of data. C4.5 is the optimal technique for decision tree generation.

$$Entropy(S) = \sum_{i=1}^n -P_i \times \log_2 P_i \quad (8)$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (9)$$

where,  $P$  is the class proportion for output,  $S$  is a set of case,  $A$  is a case attribute,  $|S_i|$  is the count of cases to  $i$ , and  $|S|$  is the count of cases in the set.

#### 4.8.3. Linear SVC

Linear SVC creates the binary classification model using the linear SVM classification model. The hyperplane of SVC generates the separation that achieves the highest distance for the nearest points of training data on any class. Hinge loss is optimized with the help of OWL query language (OWLQL) optimizer.

$$L(W; x, y) = \max(0, 1 - yW^T x) \quad (10)$$

$$f(w) = \tau R(w) + \frac{1}{n} \sum_{i=1}^n L(W; x_i, y_i) \quad (11)$$

where,  $L(W; x, y)$  is the loss function for linear SVC,  $R(w)$  is the raw tweets,  $W^T$  is the word vector,  $y$  is the sentence label and  $x$  is the emotional words for each sentence.

## 5. EXPERIMENTAL RESULTS DISCUSSION

To analyze the experimental results, the evaluation performance of the three applied classification techniques is compared to get the optimal classifier for the political sentiment analysis system on multi-class. In this system, the unstructured political dataset that contains tweets of Trump, Clinton, Obama, Joe Biden for the American presidential election and discourses in BBC News politics are used as shown in Table 1. The 80% of the total dataset is used for the training dataset and 20% of the total dataset is used for the testing dataset from each dataset. The popular performance measures (accuracy, recall, F-measure, and precision) are evaluated for this proposed analysis system.

Twitter is a social media platform that enables the posting of tweets, or brief communications, in 2016. It was used nearly 328 million active users from 2006 to 2017. During the 2016 American Presidential Election, it was one of the most widely used online social networks. A name, a profile description, a photo, and a location are the components of a Twitter user's profile. The information was gathered from August 1, 2016, to November 30, 2016, utilizing the Twitter official API to obtain user profiles and their contact networks for tweets.

Table 1. Unstructured political dataset

Dataset	Total Sentences
Trump Election (2016) Unstructured tweet	27,375
Clinton (2016)	6,444
Obama	6,852
JoeBiden	6,065
BBC NewsPolitics	1,000

For the evaluation performance of this system, the proposed lexicon-based model classifier is evaluated from the beginning of the analysis. To establish ground truth, the classified results are compared with the same data which are three machine learning classified. For improving classifier performance, data preprocessing is developed.

To analyze the performance evaluation results of this system, the political lexicon for the political domain classifier is compared with the political datasets. The experimental results of the system performance are described in Figure 2. Table 2 is shown the comparison of accuracy between each method with our proposed lexicon-based classification on a political dataset.

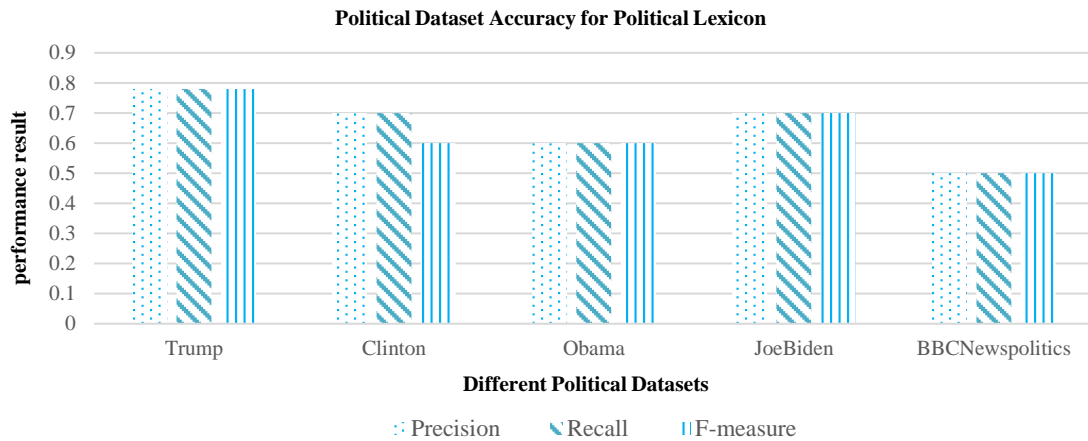


Figure 2. Percentage of tweets for political dataset on classification by political lexicon

Table 2. Political dataset accuracy for political lexicon

Dataset	Political Lexicon			Accuracy (%)
	Precision	Recall	F-measure	
Trump	0.78	0.78	0.78	93%
Clinton	0.68	0.68	0.68	89%
Obama	0.65	0.65	0.65	88.7%
JoeBiden	0.71	0.71	0.71	91.5%
BBCNewsPolitics	0.56	0.56	0.56	84.3%

According to the experimental results, strong positive, positive, neutral, negative, and strong negative have been calculated the percentage of lexicon-based classification on tweet and machine learning classification in the political dataset. The evaluation accuracy result of naïve Bayes theorem with five different political dataset is shown in Table 3. The results show that Biden dataset have the better accuracy with 71.00%.

The evaluation accuracy result of decision tree (C4.5) with five different political dataset is shown in Table 4. Biden dataset has the better performance with 99.30% accuracy in decision tree. The evaluation accuracy result of linear SVC with five political dataset is shown in Table 5. Biden dataset and BBC News politics dataset have the same accuracy of 100%.

The performance comparisons of the three classifiers on the political dataset are described in Table 6. On average, the evaluation results of precision, recall, F-measure, and accuracy of linear SVC achieve more accuracy than naïve Bayes and decision tree for political datasets. The selected model (linear SVC) is applied to classify the new collected tweets. The comparison of the machine learning classification model on the political dataset is illustrated in Figure 3.

In this section, the performance of three different models is compared. The results showed that linear SVC has the better result than the other techniques. According to the results, linear SVC which is used in one vs Rest approach classifier with proposed political lexicon-based classification model achieves the best optimal accuracy.

Table 3. Performance evaluation of political dataset with naïve Bayes

Dataset	Naïve Bayes			
	Precision	Recall	F1-Score	Accuracy
Trump	0.7460	0.6473	0.5845	64.73%
Clinton	0.7703	0.6849	0.6193	68.49%
Obama	0.7359	0.6850	0.6246	68.50%
JoeBiden	0.7991	0.71	0.6168	71.00%
BBCNewspolitics	0.7545	0.5565	0.4286	55.65%

Table 4. Performance evaluation of political dataset with decision tree (C4.5)

Dataset	Decision Tree (C4.5)			
	Precision	Recall	F1-Score	Accuracy
Trump	0.9573	0.9644	0.9621	96.44%
Clinton	0.9738	0.9668	0.9651	96.67%
Obama	0.9388	0.9348	0.9321	93.48%
JoeBiden	0.9221	0.993	0.9922	99.30%
BBCNewspolitics	0.9836	0.9838	0.9829	98.38%

Table 5. Performance evaluation of political dataset with linear SVC

Dataset	Linear SVC			
	Precision	Recall	F1-Score	Accuracy
Trump	0.9821	0.9820	0.9820	98.20%
Clinton	0.9902	0.9901	0.9901	99.01%
Obama	0.9555	0.9558	0.9554	95.58%
JoeBiden	1	1	1	100%
BBCNewspolitics	1	1	1	100%

Table 6. Overall performance evaluation of three machine learning models

Machine Learning Algorithm	Political Datasets			
	Precision	Recall	F-measure	Accuracy (%)
Multinomial Naïve Bayes	0.80	0.73	0.68	74%
Decision Tree (C4.5)	0.96	0.96	0.96	96%
Linear SVC	0.98	0.98	0.98	98%

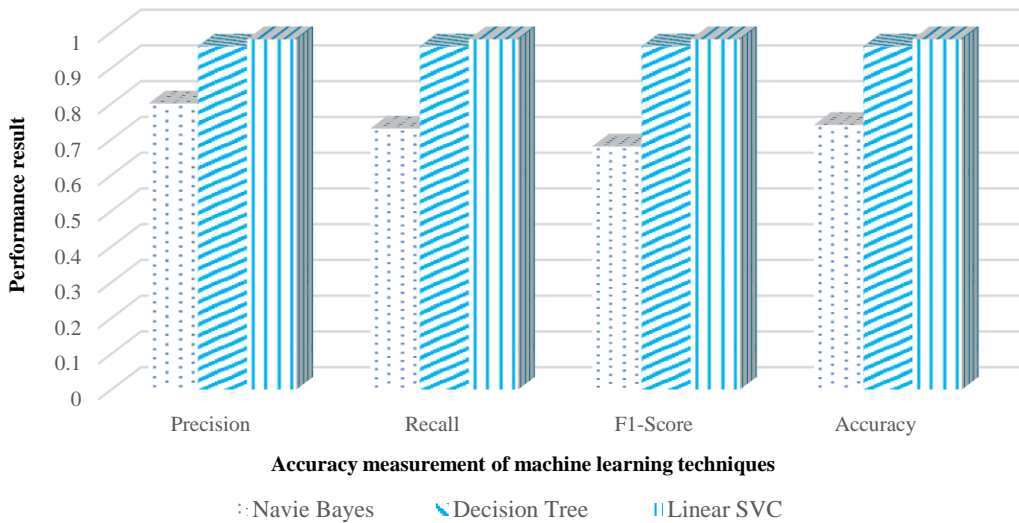


Figure 3. Performance of tweets for political dataset on classification by different model





## 6. CONCLUSION

In PMSA, system implementation is developed on the big data analytic platform (Apache Spark) for high-velocity analysis and large amounts of tweets in an effective manner. The hybrid approach is applied to get the great performance of the multi-class classification system on a vast volume of social-political data. In the proposed system, the political lexicon is created by collecting extreme opinions with their polarity, and it constructs classification model with three machine learning techniques (MNB, DT, and linear SVC). The generated lexicon is applied on different political datasets to show the performance of the lexicon. In the experimental result, the performance of political lexicon is evaluated by three difference machine learning techniques with different political datasets. According to the result, the accuracy of the linear SVC can be provided 98% that has the better performance than the other two methods. PMSA can also classify the discourse news on the political domain. In the future work, the political multi-class sentiment analysis system will also be used as a deep learning in a big data environment.





## REFERENCES

- [1] L. Vu and L. P. Thanh, "A lexicon-based method for sentiment analysis using social network data," in *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, 2017, pp. 10–16.
- [2] S. Almatarneh and P. Gamallo, "A lexicon based method to search for extreme opinions," *Plos One*, vol. 13, no. 5, May 2018, doi: 10.1371/journal.pone.0197816.
- [3] J. Feng, C. Gong, X. Li, and R. Y. K. Lau, "Automatic approach of sentiment lexicon generation for mobile shopping reviews," *Wireless Communications and Mobile Computing*, pp. 1–13, Aug. 2018, doi: 10.1155/2018/9839432.
- [4] R. Rohrschneider and F. Jung, "SS: Germany's federal election in September 2009—elections in times of duress—introduction," *Electoral Studies*, vol. 31, no. 1, pp. 1–4, 2012.
- [5] T. Fujiwara, K. Müller, and C. Schwarz, "The effect of social media on elections: evidence from the United States," National Bureau of Economic Research, Cambridge, MA, May 2021. doi: 10.3386/w28849.
- [6] M. Ringsquandl and D. Petkovic, "Analyzing political sentiment on Twitter," *2013 AAAI Spring Symposium*, 2013, pp. 40–47.
- [7] T. Elghazaly, A. Mahmoud, and H. A. Hefny, "Political sentiment analysis using twitter data," in *Proceedings of the International Conference on Internet of things and Cloud Computing*, Mar. 2016, pp. 1–5, doi: 10.1145/2896387.2896396.
- [8] J. A. Caetano, H. S. Lima, M. F. Santos, and H. T. Marques-Neto, "Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election," *Journal of Internet Services and Applications*, vol. 9, no. 1, Dec. 2018, doi: 10.1186/s13174-018-0089-0.
- [9] M. Aman Ullah, M. Arif Hasnayeem, A. Shan-A-Alahi, F. Rahman, and S. Akhter, "A search for optimal feature in political sentiment analysis," in *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, Dec. 2020, pp. 340–343, doi: 10.1109/WIECON-ECE52138.2020.9397966.
- [10] K. Ghag and K. Shah, "SentiTFIDF-sentiment classification using relative term frequency inverse document frequency," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 2, 2014, doi: 10.14569/IJACSA.2014.050206.
- [11] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, "A novel adaptable approach for sentiment analysis on big social data," *Journal of Big Data*, vol. 5, no. 1, Dec. 2018, doi: 10.1186/s40537-018-0120-0.
- [12] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of political sentiment orientations on twitter," *Procedia Computer Science*, vol. 167, pp. 1821–1828, 2020, doi: 10.1016/j.procs.2020.03.201.
- [13] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine learning-based sentiment analysis for twitter accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, Feb. 2018, doi: 10.3390/mca23010011.
- [14] A. Baltas, A. Kanavos, and A. K. Tsakalidis, "An Apache Spark implementation for sentiment analysis on twitter data," in *Algorithmic Aspects of Cloud Computing*, Springer International Publishing, 2017, pp. 15–25.
- [15] S. Wang, J. Luo, and L. Luo, "Large-scale text multiclass classification using spark ML packages," *Journal of Physics: Conference Series*, vol. 2171, 2022.
- [16] A. M. Khattak *et al.*, "Tweets classification and sentiment analysis for personalized tweets recommendation," *Complexity*, pp. 1–11, Dec. 2020, doi: 10.1155/2020/8892552.
- [17] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010, doi: 10.1002/asi.21416.
- [18] M. Bouazizi and T. Ohtsuki, "Multi-class sentiment analysis on twitter: classification performance and challenges," *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181–194, Sep. 2019, doi: 10.26599/BDMA.2019.9020002.
- [19] "Apache Hadoop," *Apache*. <https://hadoop.apache.org/> (accessed Dec. 28, 2017).
- [20] X. Meng, J. Bradley, and B. Yavuz, "MLlib: Machine learning in Apache spark," *The journal of machine learning research*, vol. 17, no. 1, pp. 1235–1241, 2016.
- [21] "Flume 1.9.0 user guide," *Apache flume*. <https://flume.apache.org/releases/content/1.9.0/FlumeUserGuide.html> (accessed Jan. 08, 2019).
- [22] "Spark Streaming," *Apache*. <https://spark.apache.org/streaming/> (accessed: Jun. 1, 2021).
- [23] "Spark MLlib," *Apache*. <https://spark.apache.org/mllib/> (accessed Jun. 01, 2021).
- [24] "ML Pipelines," *Apache*. <https://spark.apache.org/docs/3.1.2/ml-pipeline.html> (accessed May 18, 2021).
- [25] Y. M. Tun and P. H. Myint, "Comparative study for text document classification using different machine learning algorithms," *International Journal of Computer (IJC)*, vol. 33, no. 1, pp. 19–25, 2019.
- [26] A. Agrawal and T. Hamling, "Sentiment analysis of tweets to gain insights into the 2016 US election," *Columbia Undergraduate Science Journal*, vol. 11, Dec. 2021, doi: 10.52214/cusj.v11i.6359.
- [27] U. Yaqub, N. Sharma, R. Pabreja, S. A. Chun, V. Atluri, and J. Vaidya, "Location-based sentiment analyses and visualization of twitter election data," *Digital Government: Research and Practice*, vol. 1, no. 2, pp. 1–19, Apr. 2020, doi: 10.1145/3339909.

**BIOGRAPHIES OF AUTHORS**

**Yin Min Tun**     received the B.A (Geography) degree in Bachelor Arts from Mandalay University, Mandalay in 1997, Master of Computer Science in the University of Computer Studies, Mandalay (UCSM), and the M.I.Sc. degree in Computer Science from the University of Computer Studies, Yangon (UCSY) in 2000. Currently, she is an Associate Professor at the Faculty of Computer Science, University of Computer Studies (Mdy). She is currently pursuing a Ph.D. degree in Information Technology in the University of Computer Studies, Mandalay (UCSM). She can be contacted at email: yinmintun.ymtkyaw@gmail.com.



**Myo Khaing**     was born on 23<sup>rd</sup> August 1979. He received a Bachelor of Computer Science in 2003, Bachelor of Computer Science (Hons), in 2004, Master of Computer Science, in 2007, and Ph.D. (Information Technology) in January 2013. In 2005, he started working as a tutor at the University of Computer Studies, Mandalay, UCSM, and worked at the University of Computer Studies (Magway), University of Computer Studies (Maubin), served as Assistant Lecturer, Lecturer, Associate Professor, and up to the Professor. Currently, he is working as a professor at the University of Computer Studies, Mandalay, UCSM. His research interest is data mining, big data analysis, and data science. He can be contacted at email: myokhaingucsm@gmail.com.