IJEMD-CSAI, 4 (1) (2025) https://doi.org/ 10.54938/ijemdcsai.2025.04.1.418

**International Journal of Emerging Multidisciplinaries:**
**Computer Science and Artificial Intelligence**

*Research Paper*
*Journal Homepage: www.ojs.ijemd.com*
*ISSN (print): 2791-0164 ISSN (online): 2957-5036*

# Application of Support Vector Machine for Effective Prediction of Election for Sentiment Analysis

**Asoshi Paul Anule [1*], Chukwudi Jennifer Ifeoma [2], Dr. John Abiodun Oladunjoye [1]**

*1. Department of Computer Science, Federal University Wukari, Taraba State, Nigeria.*
*2. Department of ICT, Federal University Wukari, Taraba State, Nigeria.*

## Abstract

This study proposes the use of machine learning models, namely Support Vector Machine (SVM), for effective sentiment analysis on a dataset from the Kaggle repository. Considering the Tinubu 2023 election dataset, it can be seen that SVM having been fed with the cleansed dataset feature obtained an accuracy score of 93.2%, considering the result of each algorithm on the 2023 Nigerian election datasets. The study investigates data preprocessing techniques, feature selection, and correlation metrics to optimize the sentiment detection process. Results show that the SVM model achieves the highest accuracy, making it a potential tool for political analysis, business marketing, and public policy implementation. However, future research may explore deep learning techniques and data balancing strategies to enhance the models' performance further.

**Key words:** Machine learning, Opinion mining, Sentiment analysis, Support vector machine, Algorithms, Natural language processing, Neutral network, Social media.

## 1. INTRODUCTION

### 1.1 Background of the Study

The world has undergone a significant transformation due to advancements in technology and changes in sociocultural behavior. This has led to the widespread sharing of ideas, thoughts, beliefs, opinions, and decisions in real-time across various social media platforms like Twitter, Facebook, and LinkedIn [4]. However, extracting meaningful insights from this vast amount of textual data can be challenging,

**Email Address :** asoshipaulanule@gmail.com (Asoshi Paul Anule *)

primarily because of the sheer volume of information from a large user base and the difficulties associated with quantifying text data for effective modeling and decision-making [4]. Sentiment analysis, also known as opinion mining, is a field that focuses on the qualitative examination of opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards various entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It aims to categorize reviews based on predefined polarity [5]. This research area is closely linked to, or can be considered a part of, computational linguistics, natural language processing, and text mining. Building upon studies in affective state (psychology) and judgment (appraisal theory), sentiment analysis seeks to address long-standing questions in other discourse areas using data mining and computational linguistics techniques [6]. In contemporary sentiment analysis, clear and precise instructions are essential for obtaining high-quality annotations. However, sentiment reviews are often unstructured, with words having multiple meanings [7]. This ambiguity in the corpus complicates the classification of sentiment polarity. Typically, sentiment polarity for a specific text feature is divided into positive and negative categories [8]. Documents containing multiple opinionated statements often have mixed polarity, increasing the complexity of sentiment analysis [8]. It's important to note that sentiment, whether positive or negative, doesn't occur in isolation; rather than focusing on isolated compliments or complaints, sentiment analysis considers an individual's emotional expression [9]. The toxic nature of sentiment can negatively impact one's psychological well-being, potentially leading to feelings of inferiority [9]. As a result, people may become reluctant to express themselves or seek diverse opinions for decision-making. For businesses, firms, governments, and organizations, this could lead to financial losses and unmet expectations.

The significance of sentiment analysis is undeniable, as its applications and effects extend across various fields and sectors [10]. Despite its drawbacks, sentiment analysis remains a valuable tool for organizations, businesses, agencies, and governments to gain insights that facilitate efficient and effective decision-making. Given the abundance of sentiment analyzers, researchers have made numerous attempts to develop methods capable of accurately detecting sentiment from textual opinions. These applications demonstrate the viability of machine learning and deep learning models in sentiment analysis. Nevertheless, despite extensive research in sentiment analysis, current approaches continue to struggle with high false-positive rates when classifying sentiment as positive or negative. Furthermore, research on applying Machine and Deep Learning methods to sentiment analysis is still in a challenging phase, with a substantial demand for practical solutions. In response, this study proposes the use of a specific machine learning model, the Support Vector Machine (SVM), for effective sentiment detection using a dataset obtained from the Kaggle machine learning repository.

## 2. RELATED WORK

[16], Naive Bayes and Support Vector Machine (SVM) are two prominent supervised learning algorithms that play a critical role in text classification for sentiment analysis. These techniques necessitate a training dataset to learn and construct models for sentiment analysis. They are utilized to develop classifiers that sort test data into positive, negative, or neutral sentiments. To effectively implement this approach, a two-phase framework is established. The first phase involves generating training data from mined Twitter

content, which includes collecting and labeling tweets to train the sentiment analysis models. The second phase focuses on developing a scalable machine learning model to predict election outcomes based on tweet sentiments. This methodology leverages social media data and advanced machine learning techniques to provide valuable insights into public opinion and its potential impact on elections. It enables political entities and businesses to adopt a data-driven approach, making informed decisions and adjusting their strategies in response to public sentiment.

[17], The sentiment analysis classification model employs Naïve Bayes and Support Vector Machine (SVM) methods, utilizing Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction. The data analysis revealed that in March 2023, Ganjar Pranowo's data exhibited the highest positive sentiment at 77.94%, while Anies Baswedan's data showed the highest negative sentiment at 31.39%. For the November 2023 dataset, the Ganjar Pranowo - Mahfud MD candidate pair garnered the highest positive sentiment at 69.16%, whereas the Prabowo Subianto - Gibran Rakabuming Raka data displayed the highest negative sentiment at 52.12%. Frequently occurring words in positive sentiments for Ganjar Pranowo - Mahfud MD included "strong", "corruption", "support", and "appreciation". The research achieved peak accuracy rates of 86% for the SVM method and 79% for the Naive Bayes method.

[18], The research employs Naïve Bayes Classifier Algorithm and Support Vector Machine classification techniques to examine sentiment outcomes. Additionally, the investigation uses TextBlob to extract word evaluation features, categorizing text into positive or negative groups. The results show that after the text preprocessing phase of more than 15,000 tweets, 11,000 clean tweets were acquired and subsequently tagged using Python's text blob library. [1], this paper introduces a Fisher kernel function based on Probabilistic Latent Semantic Analysis for sentiment analysis using Support Vector Machine. The function is derived from the Probabilistic Latent Semantic Analysis model. This approach allows the use of latent semantic information with probability characteristics as classification features, enhancing the Support Vector Machine's classification effectiveness and addressing the issue of overlooking latent semantic features in text sentiment analysis. The findings demonstrate that the proposed method significantly outperforms comparison methods.

[19] proposed the concentrates on analyzing public sentiment regarding the 2024 presidential election. The research utilizes the SVM algorithm with Word2Vec feature extraction. The researchers chose to examine sentiment analysis of the 2024 Indonesian Presidential election using the Support Vector Machine algorithm due to its superior accuracy compared to other methods. Feature extraction is employed to enhance algorithm performance and efficiency, with Word2Vec selected for its ability to represent contextual similarities between words in generated vectors, enabling improved text classification based on context. The study's outcomes reveal optimal performance at an 80:20 ratio, achieving a precision of 88,94%, recall of 93.08%, F1-score of 90,43%, and accuracy of 90,75%. These results surpass previous research using the SVM method, which attained 82,3% accuracy.

[20] suggest examining the sentiment expressed in YouTube comments regarding three Indonesian presidential hopefuls: Anies Baswedan, Ganjar Pranowo, and Prabowo Subianto, in relation to the upcoming 2024 Presidential Election. Indonesia operates under a presidential system, with the head of

state serving a five-year term, chosen through democratic elections. The current president's term will end in 2024, prompting citizens to participate in a direct general election to select the next president for the 2024-2029 period. Leading up to the national elections, there is a significant connection to the campaign efforts of each presidential aspirant, carried out by their respective supporters. These campaign activities typically reach into rural areas and popular social media platforms, including Twitter, Facebook, and YouTube. This study seeks to evaluate the sentiment of comments on YouTube concerning three potential Indonesian presidential candidates: Anies Baswedan, Ganjar Pranowo, and Prabowo Subianto, within the framework of the forthcoming 2024 Presidential Election.

[21] aims to develop entity-level sentiment classifiers as a novel approach for forecasting the electability of presidential candidates based on citizen support on Twitter, utilizing the CRISP-DM model framework. The study evaluates 9 distinct algorithms in conjunction with 3 vectorization techniques. Performance assessment is conducted using 4 metrics: accuracy, precision, recall, and f1-score. The results indicate that TF-IDF 3-gram Random Forest achieves the highest f1-score of 0.84486. The chosen model is subsequently employed to gauge the presidential candidates' electability levels over time. In addition to simplifying the process, social media opinion mining allows candidates and their constituents to monitor electability levels cost-effectively in real-time and on-demand, offering advantages over conventional survey methods.

[2], tackled the challenge of sentiment analysis in Arabic. The authors developed a support vector machine (SVM) model to categorize Arabic micro-texts as either positive or negative. To assess the SVM model's effectiveness, they created a dataset from tweets about various social issues in Saudi Arabia, including changes related to the Saudi Arabia Vision 2030 initiative. The dataset was manually labeled based on the sentiment expressed in each text. To maximize classification accuracy, the researchers implemented several techniques within their proposed framework, such as light stemming, feature extraction (including Ngrams, emoji, and tweet-topic features), parameter optimization, and feature-set reduction. The experiments yielded impressive results, with the SVM model achieving an accuracy of 89.83%.

[3], proposed to assess the effectiveness of SVM in categorizing user feedback for the SatuSehat app into favorable and unfavorable sentiments, as well as to illustrate the most common words used in these reviews. The investigation utilized 25,000 data points, consisting of 18,359 negative and 6,641 positive class entries. The SVM classification yielded 73.4% negative sentiment and 26.6% positive sentiment. The SVM accuracy evaluation revealed a 91% overall accuracy, with positive sentiment achieving 92% precision, 71% recall, and an 80% f1-score. Negative sentiment exhibited 90% precision, 98% recall, and a 94% f1-score. Visualization showed that positive reviews frequently included words such as "good" and "great," while negative reviews often mentioned "update," "difficult," "strange," "login," and "bug." [11], The study focuses on the support vector machine technique, analyzing and implementing this approach to evaluate the outcomes. The research utilizes Weka to determine the dataset's accuracy and subsequently interpret the results.

[12], proposed to examine responses and explore the sentiment analysis process on Twitter for the TIX ID application using the support vector machine algorithm. This method is typically employed for text mining, involving data collection, data cleaning and labeling, and training and testing data distribution with three comparison scenarios: 70:30, 80:20, and 90:10, using three kernels: dot, radial, and polynomial. The process includes text preprocessing, TF-IDF word weighting, data modeling, and evaluation. The preprocessing phase comprises transform case, tokenization, and stop word filters. The findings indicate that the support vector machine algorithm achieves an accuracy of 74.17%. The study concludes that the support vector machine algorithm with 80:20 training and testing data ratio scenario produces the highest accuracy.

[13], conducted research to assess public opinion on mypertamina usage by categorizing comments using the Support Vector Machine (SVM) algorithm and identifying the most effective kernel among linear, polynomial, and RBF options. The study collected 44,400 data points from three social media platforms: 18,000 from Google Play Store, 20,000 from Twitter, and 6,400 from YouTube. Sentiment analysis was performed by assigning positive and negative classes, resulting in accuracy rates of 95% for Google Play Store, 76% for Twitter, and 99% for YouTube. The study determined that the RBF kernel outperformed linear and polynomial kernels, making it the optimal SVM kernel for this research.

[14], examined numerous Indonesian tweets related to a specific topic, classifying them using the Support Vector Machine algorithm with the Radial Basis Function kernel, employing Grid Search and cross-validation techniques. The study utilized parameters within the C and $\gamma$ ranges, aiming to achieve maximum accuracy through parameter combinations. The findings aligned with previous research, confirming that an appropriate combination of C and $\gamma$ parameters in the RBF kernel SVM method yields the highest classification accuracy.

[22] explored sentiment analysis and the efficacy of machine learning approaches in this field. Their study compared two machine learning algorithms, Random Forest (RF) and Support Vector Machine (SVM), based on their classification accuracy. The Random Forest algorithm achieved an accuracy of 0.78564, while the SVM slightly outperformed it with an accuracy of 0.80394. Both algorithms demonstrated strong performance in the classification task. The results suggest that SVM, with its marginally higher accuracy, may be preferable when accuracy is the primary concern. However, the researchers emphasize the importance of considering the specific problem requirements and basic configuration needs when selecting an algorithm for optimal results.

# 3. CONCEPTUAL FRAMEWORK

A conceptual framework delineates the methodological approach created to accomplish a given project at hand. The underlying concept of a methodology is primarily to establish a systematic framework for the successful execution of a task, to maximize efficiency and achieve desired outcomes. Therefore, it is imperative to develop a methodological framework for machine learning tasks to effectively execute the most optimal solution. Hence, the methodology employed in this study comprises three fundamental

components. In the initial phase of the study, a data preprocessing step known as the filtering technique was employed. This involved the removal of undesirable characters such as URLs and @ symbols from the dataset, and the treatment of missing values by replacing them using pandas 'fill-Na' functionality. Additionally, tokenization and lemmatization techniques were applied to separate words from the text stream where the text was then transformed into vectors through vectorization or normalization. Furthermore, the selection of features was performed using the correlation matrix, which considered the influences of dataset features. A crucial process carried out during the data preprocessing was the transformation of emoticons into their respective textural representation before tokenizing the transformed emoticons. Additionally, the dataset was divided into a standardized machine learning training set and test set in a specific proportion following the completion of the sub-stages in the data preprocessing phase.

The third methodology involves implementing two models, specifically the Support Vector Machine on the cleansed dataset. The algorithm deployed on the dataset is then used to develop a model. These model's performances were then evaluated using some performance evaluation metrics. Some of the performance evaluation metrics adapted include precision, recall, and accuracy score. The three stages of the adopted technique are illustrated in Figure: 3.1 presented below.

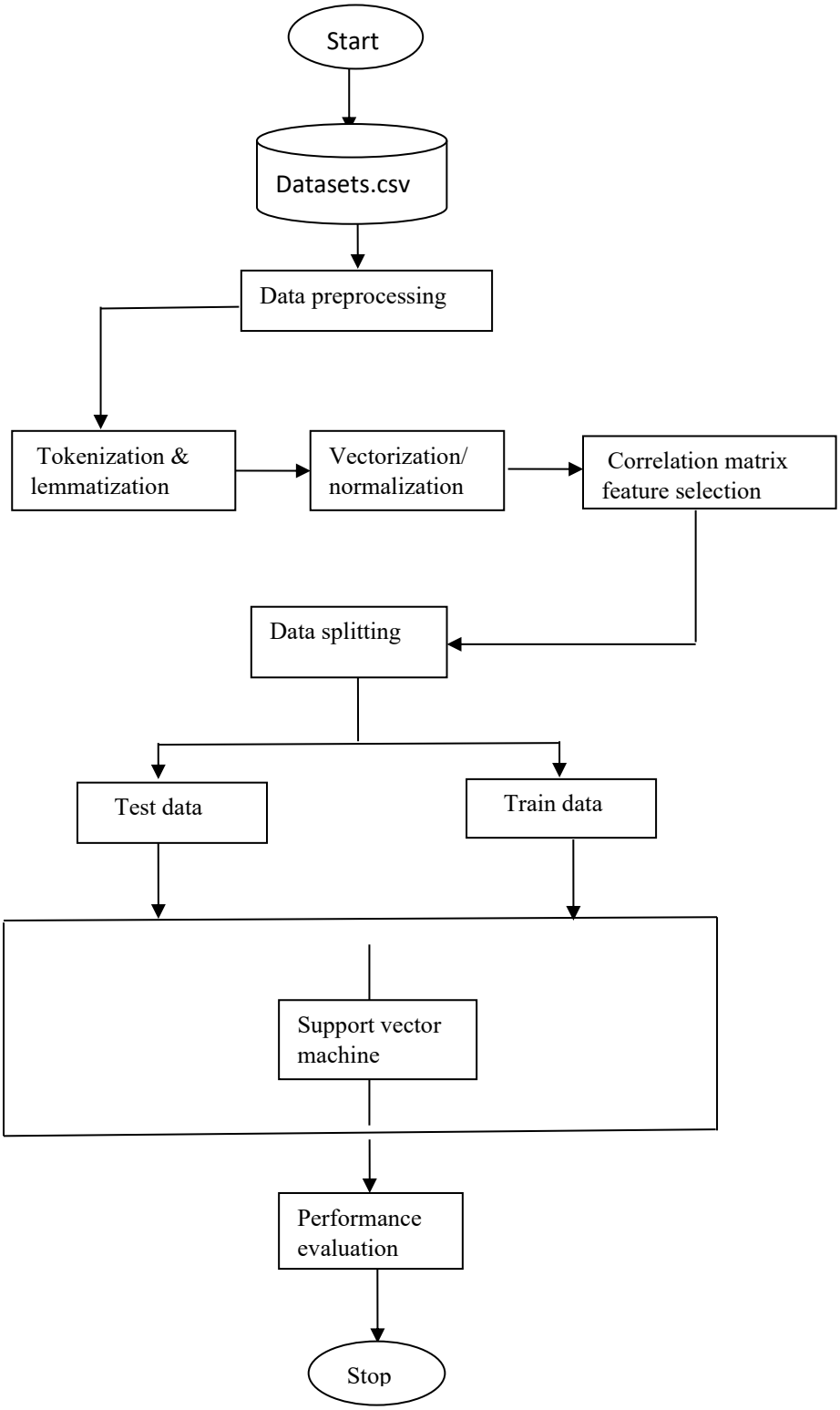


**Figure 1: Proposed Research Framework**

### 3.1 Data Source and Description

This study was carryout on two datasets sourced from the Kaggle machine learning repository website. These datasets consist of three distinct sentiment classes: Positive, Negative, and Neutral. Messages categorized as Neutral are considered non-relevant to the main entity messages; consequently, the Neutral class was treated as part of the Negative class within the datasets. The second dataset pertains to the Nigeria 2023 election and was obtained at https://www.kaggle.com/datasets/ahamefulechijoke/twitter-data-nigerian-2023-presidential-election containing nearly 10,000 tweets.

### 3.2 Data Preprocessing and Preparation

To improve model performance, it is essential to cleanse the dataset and eradicate incompleteness and inconsistency due to errors from the dataset while taking into cognizance the essential features of the dataset.

#### *3.2.1Data Preprocessing*

Data preprocessing is a critical step in sentiment analysis and machine learning in general, as it plays a pivotal role in shaping the quality and effectiveness of the models. This process involves preparing raw data for analysis by transforming and cleaning it to make it more suitable for the specific task at hand. In the context of sentiment analysis, which involves determining the sentiment or emotional tone of text data, data preprocessing serves several crucial purposes. Hence, the data preprocessing steps undertaken by this study include:

i. Removal of unwanted characters: Involves the eradication of characters such as the URLs, extra unwanted spaces, punctuation marks, and (@) symbols with the assigned user names as they are unnecessary characters that degrade the performance of the classifier.

ii. **Tokenization:** Deals with the extraction of words from a stream. In this step, the study splits the stream of text into words, phrases, or some meaningful elements into tokens through the use of delimiters such as spaces, punctuations, and symbols. The essence of tokenization is to produce different representations of information-enriched texts that can lead to better classification outcomes.

iii. **Stop words Removal:** This entails the creation of a list of words and hence scanning the document so that the word appearing in the stop list is removed.

iv. **Stemming:** is a process to reduce a word to its stem or root word. Its essence is to increase the recall rate. In addition, stemming reduces data dimensionality while identifying similar words even if they are different in shapes.

v. **Case Normalization:** Entails the conversion of upper case from the text to lowercase.

vi. **Feature Selection:** The study evaluates the correlation of the features to each other using a correlation matrix and thus selects the features that influence the target class in the prediction of sentiment.

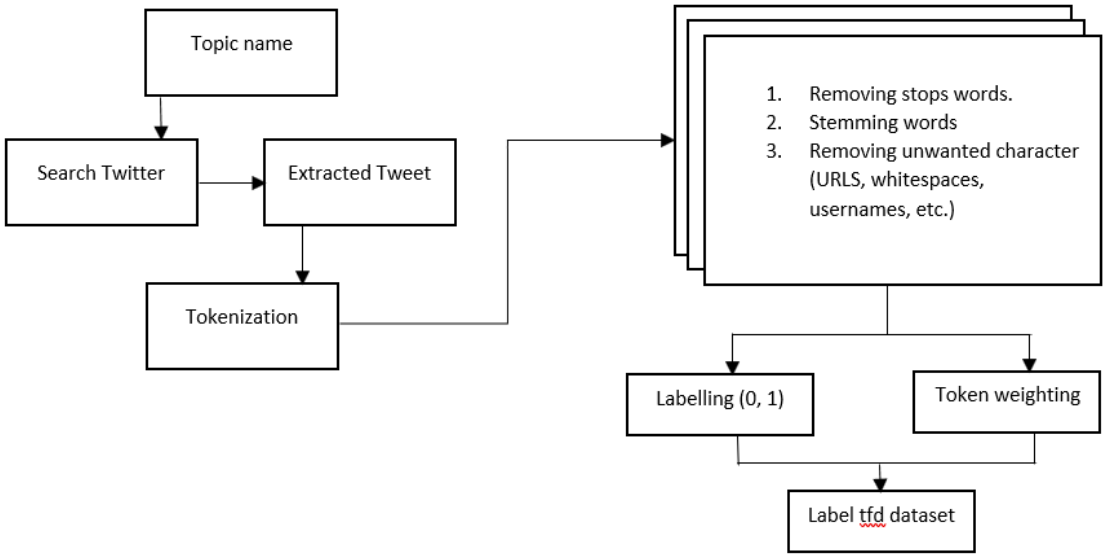


**Figure 2: Data preprocessing**

### *3.2.2 Processing of Emoticons*

In light of recent developments in the types of content observed in social media posts and messages, a new category of content has emerged, known as emoticons or emojis. Emoticons are visual representations and symbols that convey a wide range of meanings, including those associated with positive, negative, or neutral emotions. Hence, during the data preprocessing the emoticons were transformed into their corresponding textual representations for standardized analysis. The process involves conducting emoticon extraction by identifying and extracting them from the textual data considering that they are typically represented by specific combinations of characters, such as ":)" for a smiley face or ":(" for a sad face and then the creation of corpus with each corpus document representing a piece of text that contains emoticons. After this tokenization is conducted to break down the text into individual words or tokens.

## 3.3 Classification Algorithm

An extensive study conducted revealed the problem of sentiment analysis as a classification problem. A classification problem in machine learning is a task that identifies models suitable for either multiclass or a binary classification problem as in the case of this study and hence, distinguishes the classes respectively. Therefore, this study adopted classification models to evaluate their performance based on sentiment analysis via distinct evaluation metrics. This model is the Support Vector Machine.

### *3.3.1 Support Vector Machine*

The Support Vector Machine (SVM) classifier is one of the most popular models applicable in a wide range of classification tasks, especially natural language classification problems. Although, the Support Vector Machines is considered to be a classification approach but can be applied in both classification and regression problems as it can handle multiple continuous and categorical variables. The SVM model

constructs a hyper plane iteratively in multidimensional space to separate distinct classes of a problem and thus generates an optimal hyper plane to minimize or cutdown the error rate. From Figure 3.2, the support vectors define the data points, which are closest to the hyperplane using a Linear variation of the Support Vector Machine. These points define the separating line better by calculating margins that are more relevant to the construction of the classifier. In the scenario where the margin is larger in between the classes, then it is considered a good margin while a smaller margin is a bad margin.

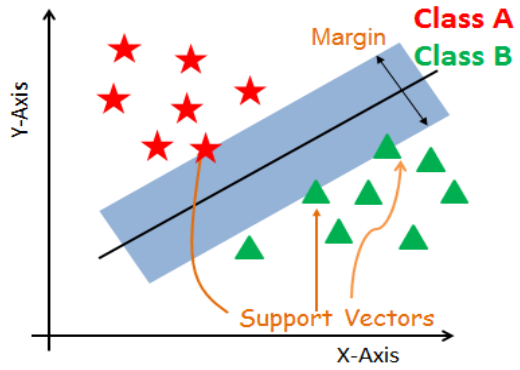


**Figure 3: Support Vector Machine**

**Source**:(https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python)

**Algorithm 1:** ***Support Vector Machine Algorithm***

*Step 1:* Simple SVM candidateSV = {closest pair from opposite classes}

*Step 2:* initialize*i*

*Step 3:* while there are violating points do

*Step 4:* Find a violator

*Step 5:* candidate SV = candidate SVx violator

*Step 6:* if any $\partial_i < 0$ due to addition of c to S then

*Step 7:* candidate SV = candidate SV/*i*

*Step 8:* repeat till all such points are pruned

*Step 9:* end if

*Step 10:* increment*i*

*Step 11:* end while

Algorithm 3.1 outlines the steps for a Support Vector Machine (SVM), a powerful machine-learning technique used for classification tasks. SVMs excel at finding an optimal decision boundary that maximizes the margin between different classes in the dataset. This algorithm appears to focus on the process of selecting support vectors, which are crucial data points used to define the decision boundary.

Step 1: The algorithm starts by initializing a set called candidateSV with the closest pair of data points from opposite classes. These points are chosen because they are critical for defining the margin between the classes, which is one of the key principles of SVMs.

Step 2: Initialize a variable i. The purpose of this variable will become clear in the subsequent steps.

Steps 3-11: These steps form a loop that continues as long as there are violating data points. Violating points are data points that fall on the wrong side of the decision boundary or within the margin.

Step 4: Within the loop, the algorithm finds a violator, which is a data point that violates the margin constraints. This means the point is either misclassified or too close to the decision boundary.

Step 5: The violator is added to the candidateSV set. This set now contains additional support vectors.

Step 6: If any support vector has a negative partial derivative $\partial_i$ due to the addition of the candidate violator, suggests that this support vector might no longer be necessary for defining the margin. Therefore, the candidate violator is divided by i, which is gradually incremented (Step 10). This process is repeated

until all such points are pruned (Step 8). The intention is to optimize the set of support vectors to improve the SVM's efficiency.

### 3.4 Performance metrics

To evaluate the performance of the two adopted classification models namely, the Support Vector Machine and the Linear Regression on the sourced sentiment dataset. This study utilized the following evaluation metrics:

**Confusion Matrix:** The confusion matrix although not an accuracy metric is used to determine each model's correctness and accuracy. It, therefore, serves as a checker for the utilized accuracy score. The confusion matrix validates the accuracy parameter using the following indicators as shown in Table 3.1 below.

**Table 1: Confusion Matrix**

| | | **Actual** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Predicted** | Positive | TP | FP |
| | Negative | FN | TN |

i. **True Positives (TP):** Defines a scenario where a data point is initially True and the model reviewed and anticipated it True.

ii. **True Negatives (TN):** Defines a scenario where a data point is initially False and the model reviewed and anticipated it to be False.

iii. **False Positives (FP):** Indicates an instance where the actual class data value was False but the model anticipated true.

iv. **False Negatives (FN):** Indicates an instance where the actual class data value was True but the model anticipated False.

**The precision:** Accuracy defines the percentage of the number of correctly predicted positive reviews divided by the total number of predicted positive reviews:

$$precision = \frac{TP}{TP + FP} \tag{6}$$

**Recall:** defines the classifier's completeness. It is said to be the percentage of correctly predicted positive reviews to the actual number of positive reviews from the dataset. Thus, recall can be mathematically written as:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

**F-score:** harmonize the metrics score obtained by the precision and recall metrics to achieve the best score value (1) or the worst scores value (0):

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

**Accuracy:** of all metrics, the accuracy metric is the most significant performance evaluation variable. It is defined as the percentage of the number of correctly predicted reviews to the total number of reviews present in the dataset and is mathematically represented as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

# 4. RESULT AND DISCUSSIONS

## 4.1 Introduction

This chapter presents the result of the machine learning classification algorithms namely, the Support Vector Machine which was utilized in the classification of sentiment tweets from the 2023 Nigerian election dataset sourced from the Kaggle machine learning repository. The result of the model is presented in tabular form. The score of each algorithm is validated against some performance evaluation metrics such as the precision, recall, confusion matrix and f1-score. Before the presentation of the result data visualization and exploration were conducted to understand the correlation between the dataset features with a graphical visualization of the dataset feature using the graph plotter from the seaborn and mat-plot libraries.

## 4.2 Dataset Description and Visualization

Prior to data exploration and visualization, the integration of the adapted dataset is an essential phase that promotes the exploration and visualization of the patterns in the identification of sentiment in tweets. Hence, in reading the dataset, Panda's framework was utilized as it provides distinct functions for the importation of the datasets of distinct file formats into machine learning project codes. For reading the 2023 Nigerian election datasets into the program the panda's 'read_csv' function enables the readability of the datasets considering that both datasets have commas-separated value file format extensions. The code snippets in Figure 4.1 provide insight into reading the datasets and require only changing the file name during the course of reading the dataset.

```
In [*]: df =pd.read_csv('nigeria_election.csv', encoding =DATASET_ENCODING)
```

```
In [4]: df.head()
```

Out[4]:

| | id | user_name | user_location | user_description | user_created | user_followers | user_friends | user_favourites | user_verified | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1617619263392743424 | Fabulous Faga | Abakaliki, Nigeria | ð¸ð¹A Social media influencer\nð¸ð¹ A scientist... | 2020-10-19 08:02:47+00:00 | 244.0 | 999.0 | 686.0 | False | 2023-0 20:24:20+0 |
| 1 | 1617648566977302534 | Patriotic Nigerian | Lagos, Nigeria | inspiring and amazingly created. Marketing/ Br... | 2021-03-30 22:04:11+00:00 | 688.0 | 2827.0 | 3488.0 | False | 2023-0 22:20:47+0 |
| 2 | 1617657018873171969 | Promise | NaN | Health and Fitness enthusiast | 2022-05-06 23:03:08+00:00 | 2.0 | 22.0 | 40.0 | False | 2023-0 22:54:22+0 |
| 3 | 1617584206317752321 | KemKem | Lagos, Nigeria | My Tweets & Views are Personal & do not\nrefle... | 2015-01-01 07:40:09+00:00 | 24474.0 | 1931.0 | 129947.0 | False | 2023-0 18:05:02+0 |
| 4 | 1617619489184690178 | Mo'6icky | NaN | By nature, an optimist_ð¸ð¹ð¦ | 2019-05-09 16:31:33+00:00 | 22.0 | 198.0 | 540.0 | False | 2023-0 20:25:14+0 |

**Figure 4: Reading the Dataset into the Program**

Figure 4.2 shows the data type of each feature set attribute from 2023 Nigerian election datasets. It can be seen that the dataset features vary with different types and hence requires the conversion of each feature type to the same data type before feeding it to the models. The conversion of the data type to a numeric type constitutes part of the data preprocessing part that ensures the dimension of each feature is in numeric format.

```
In [22]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                5000 non-null   object
 1   user_name         5000 non-null   object
 2   user_location     3967 non-null   object
 3   user_description  4705 non-null   object
 4   user_created      5000 non-null   object
 5   user_followers    5000 non-null   object
 6   user_friends      5000 non-null   object
 7   user_favourites   5000 non-null   object
 8   user_verified     5000 non-null   object
 9   date              5000 non-null   object
 10  text              5000 non-null   object
 11  hashtags          2226 non-null   object
 12  source            5000 non-null   object
 13  retweets          5000 non-null   object
 14  favorites         5000 non-null   float64
 15  is_retweet        5000 non-null   object
dtypes: float64(1), object(15)
memory usage: 625.1+ KB
```

**Figure 5: 2023 Nigerian Feature Data Types**

The above Figures shows the analysis of the number of positive sentiments in proportion to words that have negative sentiment polarity for the 2023 Nigerian election datasets respectively. In detail, where as for the Nigerian 2023 election, the records shows that 1184 has a negative polarity while 8816 has a

positive polarity of the total 10,000. From the figures on 4.2 (a, b), 0 represent the negative and 1 defines the positive polarity.

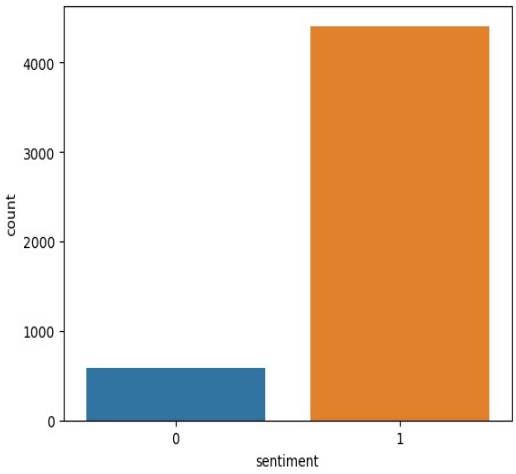


**Figure 6: Sentiment Count (Nigerian).**

The understanding of data correlation is essential in identifying features that greatly influence the prediction of sentiments with positive and negative polarities. The correlation matrix in Figure 4.3, uses the correlation coefficients annotated to each cell to establish the degree to which each feature is connected in the prediction of sentiment polarity as either positive or negative. From the correlation diagram, a value of 0 denotes a neutral correlation, a value of -1 denotes a weak correlation, and a value of 1 denotes a substantial influence between two factors in predicting cases whether or not an instance has negative or positive sentiment polarity. The diagonal axis is always equal to one because each attribute has a strong association with itself. Furthermore, in view of the right-hand side of the data correlation diagram in Figure 4.4, the top right-hand-side value identifies features that significantly influence, the prediction of sentiments with a higher value closer to 1. Essentially, it is important to note that the lower the correlation values, the least significant a feature is to the prediction of sentiments. Moreover, the degree of light orange shading of a feature determines the degree of a feature's correlation with other features in influencing the prediction of sentiments with the reddish and blackish shaded cells indicating the least correlation of its respective horizontal and vertical cells features.

Figure 4.5 and 4.6 shows the word count diagram of the 2016 Donald Trump and Nigerian 2023 election dataset respectively. The word counts diagram describes the frequently used word for the given text corpus from the datasets. It is used to preprocess the raw text data by converting the text into numerical representation while assigning some frequencies to each word.

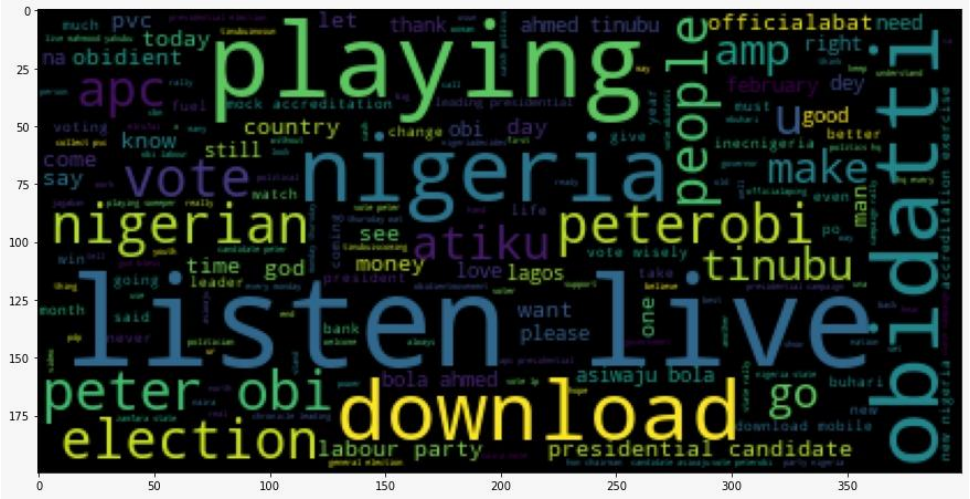


**Figure 7: Nigerian 2023 Election Word Count**

### 4.3 Data Scaling

In any machine learning experiment, the normalization of the dataset during the cost of data preprocessing is an essential aspect that is needed to enhance the performance of a model. Moreover, it is essential to scale the dataset into a feature range, enhancing the model's capabilities in the classification of sentiment as positive or negative. To ensure feature scaling this study utilized the standard scaler functionality to scale the dataset features into a feature ranging value of zero and one. An ideal fact about scaling the dataset features is to enhance the model's quick convergence with a confined range of variables.

```
In [81]: # features scaling
         scaler = StandardScaler()
         x[x.columns]= scaler.fit_transform(x)
```

**Figure 8:** Dataset Scaling

```
In [12]: le =  LabelEncoder()
         for i in df:
             if df[i].dtype=='object':
                 df[i] = le.fit_transform(df[i])
             else:
                 continue
```

**Figure 9: Dataset Encoding**

### 4.4 Percentage Split Technique

Given that splitting the dataset into some training and test proportions is necessary for training machine learning experiments. The experiment also splits both the 2023 Nigerian election datasets into some training and test proportions. SVM algorithms were trained using the training set, and the test set was used to verify the effectiveness of the produced models (as a result of the training process). To be more specific, during the data splitting stage, 70% of the dataset was used to train the models, and the remaining 30%

for each dataset was used to test the efficacy of the developed models following best practices in machine learning.

### 4.5 Result Presentation

Having explored the dataset feature (backup with data preprocessing). It was identified that the dataset suits the classification problem considering that the polarity of sentiment from the dataset ranges between 0 and 1 identifying a binary classification problem. Hence, the study as aforementioned harness the classification potential of three machine learning algorithms namely the SVM classifier for the prediction of sentiment polarity (as 0 and 1 that is positive or negative) on the cleansed and scaled features. The accuracy of each algorithm used is presented in Table below representing the 2023 Nigerian election dataset respectively. It can be seen that SVM having been fed with the cleansed dataset feature obtained an accuracy score of 93.2

**Table 2: Accuracy Score (Nigeria 2023 election)**

| **Algorithm** | **Accuracy (%)** |
|---|---|
| Support Vector Machine | 93.2 |

### 4.6 Evaluation Metrics

To evaluate the performance accuracies as obtained by each algorithm, the study as aforementioned proposed the evaluation of each model's accuracy in correspondence to the score obtained by each algorithm using the precision, recall and f1-score metrics to rigorously analyze their performances. Table 4.3 show the evaluation metrics scores for each of the model on the Donald Trump 2016 election dataset while Table 4.4 depicts the accuracy metrics for the 2023 Nigerian election.

**Table 3: Evaluation metrics (2023 Nigerian election)**

| **Algorithm** | **Precision (%)** | **Recall (%)** | **F1-score (%)** | **Accuracy** |
|---|---|---|---|---|
| SVM | 93 | 97 | 95 | 93.2 |

### 4.7 Confusion Matrix

To extensively evaluate the performance of the SVM the confusion matrix was also employed. Hence, taking into cognizance the confusion diagram shows the x and y-axis represent the class label for the actual and predicted respectively with their prediction scores annotated to the intersecting cells. The first quadrant of the confusion matrix represents the True Positive (TP) which represents a scenario where the predicted value is positive, and the instance model (SVM) predicted it as positive. The second quadrant of the confusion matrix diagram represents the False Positive (FP) metric and identifies the condition where the model predicted value is positive, but it is false from the adapted sentiment dataset label. The

third defines the False Negative (FN) metric and thus identifies a scenario where the developed model's predicted value is negative, but it is initially positive from the test data. And lastly, the fourth quadrant which is identified as the True Negative (TN), identifies cases where the predicted value is negative and is negative from the test data.

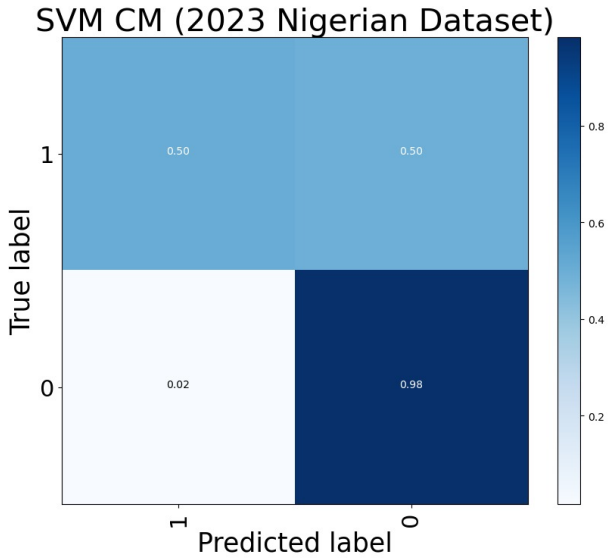


**Figure 10: SVM CM Nigerian 2023 Election**

# 5. CONTRIBUTION OF KNOWLEDGE

The key application area of this implementation is in the political discipline. Hence, political election analysts can utilize the viabilities of the implemented model to analyze sentiment during the election period as to the emergence of the election result.

The algorithm utilized can be applied in other natural language processing fields related to sentiment analysis also such as business sentiment analysis, crypto market sentiment analysis, etc.

The computation concept and algorithm of the designed model implementation can be utilized in various institutions to expose students to the coherent complexities of computational models and their respective efficiency and effectiveness.

# 6. CONCLUSION

In this study, models for the classification of sentiment polarities from the election dataset from the Nigerian datasets sourced from the Kaggle machine learning repository have been experimentally evaluated and analyzed using machine learning algorithms such as the SVM, achieved an accuracy of 93.2.

# 7. REFERENCES


[1] Han, K. X., Chien, W., Chiu, C. C., & Cheng, Y. T. (2020). Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. *Applied Sciences*, *10*(3), 1125; https://doi.org/10.3390/app10031125

[2] Alyami, S. N., & Olatunji, S. O. (2020). Application of support vector machine for Arabic sentiment classification using twitter-based dataset. *Journal of Information & Knowledge Management*, *19*(01), 2040018. https://doi.org/10.1142/S0219649220400183

[3] Imanuddin, S. H., Adi, K., & Gernowo, R. (2023). Sentiment Analysis on Satusehat Application Using Support Vector Machine Method. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *5*(3), 143-149. https://doi.org/10.35882/jeemi.v5i3.304

[4] Saif, H., Fernandez, M., Kastler, L., & Alani, H. (2017). Sentiment lexicon adaptation with context and semantics for the social web. *Semantic Web*, *8*(5), 643-665. 10.3233/SW-170265

[5] Dey, R. K., Sarddar, D., Sarkar, I., Bose, R., & Roy, S. (2020). A literature survey on sentiment analysis techniques involving social media and online platforms. *International Journal Of Scientific & Technology Research*, *1*(1), 166-173.

[6] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

[7] Agaian, S., & Kolm, P. (2017). Financial sentiment analysis using machine learning techniques. *International Journal of Investment Management and Financial Innovations*, *3*(1), 1-9.

[8] Phan, H. T., Tran, V. C., Nguyen, N. T., & Hwang, D. (2020). Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. *Ieee Access*, *8*, 14630-14641. 10.1109/ACCESS.2019.2963702

[9] Cachola, I., Holgate, E., Preoţiuc-Pietro, D., & Li, J. J. (2018). Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2927-2938).

[10] Hasan, M., Rundensteiner, E., & Agu, E. (2019). Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, *7*, 35-51. https://doi.org/10.1007/s41060-018-0096-z

[11] Gupta, A., Tyagi, P., Choudhury, T., & Shamoon, M. (2019). Sentiment analysis using support vector machine. In *2019 International conference on contemporary computing and informatics (IC3I)* (pp. 49-53). IEEE. 10.1007/s41060-018-0096-z

[12] Nabillah, A., Alam, S., & Resmi, M. G. (2022). Twitter User Sentiment Analysis Of TIX ID Applications Using Support Vector Machine Algorithm. *RISTEC: Research in Information Systems and Technology*, *3*(1), 14-27.

[13] Nursalim, A., & Novita, R. (2023). Sentiment Analysis of Comments on Google Play Store, Twitter and Youtube To the Mypertamina Application With Support Vector Machine. *Jurnal Teknik Informatika (JUTIF)*, *4*(6), 1305-1312. https://doi.org/10.52436

[14] Ferdiansyah, H., Komaria, N., & Arief, I. (2023). The Application of Support Vector Machine Method to Analyze the Sentiments of Netizens on Social Media Regarding the Accessibility of Disabilities in Public Spaces. *Journal of Information System, Technology and Engineering*, *1*(1), 6-10. https://doi.org/10.61487/jiste.v1i1.8

[15] Khan, T. A., Sadiq, R. ., Shahid, Z. ., Alam, M. M., & Mohd Su'ud, M. B. . (2024). Sentiment Analysis using Support Vector Machine and Random Forest. *Journal of Informatics and Web Engineering*, *3*(1), 67–75. https://doi.org/10.33093/jiwe.2024.3.1.5

[16] Gupta, S., Gaur, S. S., Sharmaa, P. R. A. T. I. B. H. A., & Gupta, A. (2024). Election Prediction Using Twitter Sentiment Analysis Using Naïve Bayes and Support Vector Machine. *Available at SSRN 4754965*. http://dx.doi.org/10.2139/ssrn.4754965

[17] Firdaus, A. A., Yudhana, A., & Riadi, I. (2024). Prediction of Presidential Election Results using Sentiment Analysis with Pre and Post Candidate Registration Data. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, *10*(1), 36-46. https://doi.org/10.23917/khif.v10i1.4836

[18] Dharta, F. Y., Mahardhani, A. J., Yahya, S. R., Dirsa, A., & Usulu, E. M. (2024). Application of Naive Bayes Classifier Method to Analyze Social Media User Sentiment Towards the Presidential Election Phase. *Jurnal Informasi dan Teknologi*, 176-181. https://doi.org/10.60083/jidt.v6i1.494

[19] Damayanti, L., & Lhaksmana, K. M. (2024). Sentiment analysis of the 2024 Indonesia presidential election on Twitter. *Sinkron: jurnal dan penelitian teknik informatika*, *8*(2), 938-946.10.33395/sinkron.v8i2.13379

[20] Wahyudi, A., Santoso, G. B., & Sholihah, B. (2024). Analysis of The Sentiment of Indonesian Presidential Candidates for 2024 on The YouTube Social Media Platform using The Support Vector Machine Method. *Intelmatics*, *4*(1), 22-30. https://doi.org/10.25105/itm.v4i1.17636

[21] Fauzi, A., Butar, J. B., Budi, I., Ramadiah, A., Putra, P. K., & Santoso, A. B. (2024). Supervised Machine Learning Entity Sentiment Analysis: Prediction of Support for 2024 Indonesian Presidential Candidates. *Revue d'Intelligence Artificielle*, *38*(2). **10.18280/ria.380222**

[22] Khan, T. A., Sadiq, R., Shahid, Z., Alam, M. M., & Su'ud, M. B. M. (2024). Sentiment Analysis using Support Vector Machine and Random Forest. *Journal of Informatics and Web Engineering*, *3*(1), 67-75. https://doi.org/10.33093/jiwe.2024.3.1.5