Sentiment Analysis of Political Tweets in the Philippine Election 2022 using Machine Learning

Kyle A. Destura
Bicol University
Sorsogon City, Sorsogon
0910 746 5182
kyleafundar.destura@bicol-u.edu.ph

Eden Rose Nate
Bicol University
Comun, Camalig, Albay
0912 938 5083
edenrosemagallon.nate@bicol-u.edu.ph

China Jepsane
Bicol University
San Isidro Castilla, Sorsogon
0915 260 5010
chinasolitas.jepsane@bicol-u.edu.ph

Ralph Robin Ponta-oy
Bicol University
Bulacao, Gubat, Sorsogon
0928 588 1209
ralphrobinespectacion.ponta-oy@bicol-u.edu.ph

ABSTRACT

Twitter has become the most popular social media platform for expressing one's thoughts and opinions. Emerging events or news are frequently accompanied by an increase in Twitter activity, providing a unique opportunity to assess the relationship between expressed public sentiment and electoral outcomes. The exploding twitter data gathered as a result of the political campaign can be utilized to predict the presidential election, as it has been done in other countries. Sentiment analysis is related to the problem of extracting sentiments from publicly available data. The main aim of the researchers is to present an analytical study using Twitter text datasets to interpret public opinion towards the candidates. In this study, Twitter data was collected using TWINT from the period of February 4 to March 28. The collected tweets were manually annotated and R is used for preprocessing and analyzing the tweets using machine learning approaches. Classifying the tweets into positive, neutral and negative sentiments was also the main objective of this study. Furthermore, identifying and analyzing the public sentiments toward the three most popular presidential candidates in order to determine their prospects of being elected into the Philippines' highest position of power based on tweets.

Keywords

Twitter, TWINT, Machine Learning, Sentiment Analysis, Classification, Decision Trees, SVM, Naïve Bayes, Random Forest

1. INTRODUCTION

As the world continues to develop, innovate, and change, it came across with the creation of "Sentiment Analysis" in the 1950s. Currently at the present, it is widely used in businesses to assess customer satisfaction and reviews, product and brand recognition, and product advertisement and promotion. Sentiment analysis is a method of determining consumer likes, dislikes, comments, opinions, or feedback about content that will be categorized as positive, negative, or neutral. In terms of sentiment analysis, social media plays a significant role. According to the Digital 2022: The Philippines by Simon Kemp, the Philippines' total population was 111.8 million in January 2022. There were 92.05 million social media users in the Philippines in January 2022[1].

Twitter has recently emerged as the preferred social media platform for obtaining datasets. Large unstructured data sets have become more accessible to everyone due to an increase in the number of users worldwide. People can efficiently collect data and information to organize and manipulate for a specific purpose, which is evident in sentiment analysis. In this regard, election result forecasting is an application of sentiment analysis aimed at predicting the outcomes of an ongoing election by gauging public opinions via social media [2].

Twitter is a text-based micro-blogging and social networking service. Tweets are the messages that are shared on this social media network. Twitter allows users to convey their thoughts in a more precise and meaningful way. As a result, this study focuses on Twitter data in order to provide more reliable data for sentiment analysis as part of the classification method. With this, this study proposed to classify Twitter text datasets to determine the people's sentiments towards the Philippine Election, with a focus on the Top 3 Presidential candidates in the 2022 election. The collected datasets were manually annotated and analyzed based on the results of machine learning approaches. This study also aims to present a data representation of the sentiment results and to compare the performance of machine learning approaches in order to identify the best classification model for Twitter text datasets.

2. RELATED LITERATURE

2.1 Election-Related Tweets as Data

Java et. al.[3] People use microblogging to talk about their daily activities and to seek or share information. Twitter is a very popular social medium and it contains over millions of Filipino users. It is being used as a way to communicate experiences, views and ideas of an individual via tweets. Twitter users can provide hashtags to their tweets which are likely connected to the topic. The researchers have utilized the hashtags to extract tweets concerning sentiments around a particular presidential candidate.

Twitter has recently emerged as the preferred social media platform for obtaining datasets. Large unstructured data sets have become more accessible to everyone due to an increase in the number of users worldwide. People can efficiently collect data and information to organize and manipulate for a specific purpose,

which is evident in sentiment analysis. In this regard, election result forecasting is an application of sentiment analysis aimed at predicting the outcomes of an ongoing election by gauging public opinions via social media [2].

According to Cozma and Chen [4], sentiment analysis on Twitter is a new and difficult field; with the growing number of online users, the quality of text has become a major concern for observers and analysts. The increased amount of users implies the increased amount of informal texts such typographical errors, grammatical errors, shortcut texts, repeated letters and etc. Filipinos are known for their informal texts, even creating their own style of texting such as the infamous jejemon. This proves that sentiment analysis is a broad study in the field of Natural Language Processing.

2.2 Preprocessing of Data

Because of the size of the datasets used for data mining, the data preprocessing step has become so important that it has been designated as a data mining technique. Data processing techniques, when used prior to mining, can significantly improve the quality of the patterns mined and/or the time required for the actual mining. The following are the fundamental data preprocessing techniques are data cleaning, data transformation, data reduction and discretization. Filling in missing values, smoothing out noise, dealing with outliers, and detecting and removing redundant data are all part of data cleaning. When necessary, data transformation converts the data into appropriate forms for mining. Data reduction is used to reduce the size of the data set to be mined. While the 'dimension reduction' technique removes superfluous attributes, the 'data compression' and 'numerosity reduction' techniques provide alternative forms of reduced data representations. Lastly, discretization is a type of data reduction in which low-level concepts are collected and replaced with high-level concepts to reduce the number of levels of an attribute [5].

Dařena and Žižka [6] evaluated existing non-standard short text preprocessing methods and applications, and revealed numerous informative patterns. They claim smaller datasets are ineffective and yield inefficient results. The quality of data gathered have tremendous effect on the results of the data cleaning. The researchers must have a good model for filtering out data when extracting tweets from twitter. Some hashtag sentiments are biased, and this must be taken into account for a more valid data set. Text normalization is not just a single step, it is the process of systematically conducting a sequence of essential actions. i.e. Tokenization, Stop word removal, Part of Speech Tagging, Stemming and Lemmatization. This implies that text cleaning or data cleaning is not an objective process. To be able to deliver reliable conclusions from a study, researchers must have a thorough understanding of the targeted people's language.

2.3 Machine Learning Approaches

According to the study of Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar and Mohammad Masum [7], they used SVC linear kernel for the Support Vector Machine and for the Naïve Bayes, multinomial Naïve Bayes has been used. The *Naive Bayes* classifier is a set of Bayes Theorem-based classification algorithms which is utilized as a probabilistic classifier. When used for text data analysis, the Naive Bayes classifier produces great results where Natural Language Processing as an example. The Support Vector Machine was also applied wherein the input is vector space, and the output is either positive or negative. The result of the experiment gave 83% accuracy for SVM while the Naïve Bayes yielded 77%. The support vector machine (SVM) has been demonstrated to be useful in sentiment analysis. SVM examines information, categorizes choice limits, and employs components in the input space for calculation. The critical data is presented in two vector configurations. Each datum (expressed as a vector) is now assigned to a class of size m. The machine then finds the boundary between the two classes that does not appear in any of the training samples. The distinct distinguishes the classification edge, and broadening the edge reduces ambiguity choices. The SVM has been shown to outperform the Nave Bayes classifier in a variety of text classification problems [8].

Sentiment analysis may be used to calculate, recognize, and communicate emotion in a variety of fields. One of the best classification methods is the *Random Forest Algorithm*. It is capable of accurately classifying large volumes of data. It is a classification and regression ensemble learning approach that builds a number of decision trees during training and offers the class that is the mode of the classes generated by individual trees. The *Support Vector Machine* (SVM) was also applied in which it's a supervised machine learning technique that may be applied to both classification and regression problems. Based on the study of Yassine Al Amrani, Mohamed Lazaar, and Kamal Eddine El Kadiri [9], the accuracy, recall, and F-measure for RFSVM were 83.4 percent, 83.4 percent, and 83.4 percent, respectively. On the criteria of Accuracy, Precision, Recall, and FMeasure, the hybrid approach, which combines Random Forest and Support Vector Machine, produced better results. As a single hybrid approach, the Random Forest approach increased performance in the case of smaller reviews while the Support Vector Machine approach improved performance in the case of larger reviews.

3. METHODOLOGY

The purpose of this paper is to classify public tweets toward the three most popular presidential candidates in this 2022 election which includes Ferdinand "Bongbong" Marcos Jr., Isko Moreno Domagoso and Leni Robredo and evaluate the performance of machine learning approaches. A range of different of processes were also included in order to deeply understand the flow of methods required for this study.

The following are the steps taken throughout the methodology.

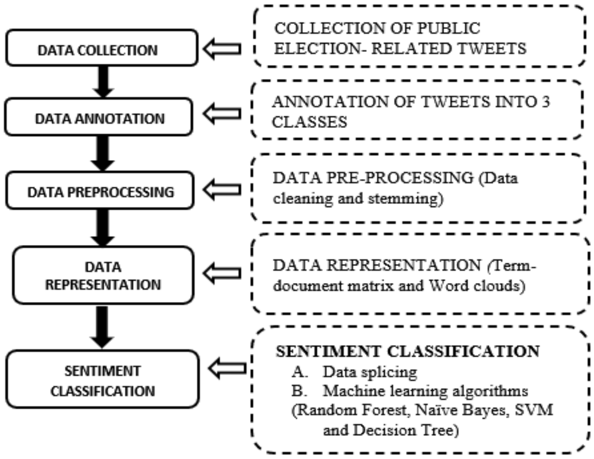

DATA COLLECTION
COLLECTION OF PUBLIC ELECTION- RELATED TWEETS
DATA ANNOTATION
ANNOTATION OF TWEETS INTO 3 CLASSES
DATA PREPROCESSING
DATA PRE-PROCESSING (Data cleaning and stemming)
DATA REPRESENTATION
DATA REPRESENTATION (Term-document matrix and Word clouds)
SENTIMENT CLASSIFICATION
SENTIMENT CLASSIFICATION
A. Data splicing
B. Machine learning algorithms (Random Forest, Naïve Bayes, SVM and Decision Tree)

Figure 1. Proposed Methodology

3.1 Data Collection

The researchers took advantage of a public API 'TWINT'. It has been prioritized over the Twitter API because of the reasons of low privileges provided by it. TWINT allowed the researchers to bypass the constraints of the Twitter API, it also has the advantage of being easy to use, researchers with little experience of python coding, javascript, and Twitter API may use it to extract tweets. Table 1 reveals the number of tweets extracted.

Table 1. Tweets Collected

Candidate	Tweets
Maria Leonor "Leni" G. Robredo	27,513
Ferdinand “Bongbong” R. Marcos Jr.	11,118
Francisco “Isko” Moreno Domagoso	3,565
Total	42,196

3.1.1 Issues encountered on Data Collection

3.1.1.1. Presidential Candidates’ Popularity.

Presidential candidates may be more popular in twitter but not in other social media platform such as in the case of candidate Ferdinand Marcos Jr’s popularity vs. Leni Robredo’s popularity in twitter, it is shown that Robredo is a more popular candidate on twitter based on the hashtags and amount of tweets extracted while Marcos is a more candidate in Meta’s Facebook platform. Popularity may result to imbalanced dataset and overall result of the study. Different social medium must be considered to provide a more valid sentimental analysis result.

3.1.1.2 Hashtag Trends

It has been observed by the researchers that hashtag do age. Hashtags also reflects events such as ‘#PasigLaban’, ‘#SolidNorth’, ‘#SMNIdebate’, in which after a week or two of the event, it likely that a hashtag is no longer used. The date in which these hashtag is used as a filter results to more data gathered resulting to a skewed dataset. This issue implies that the researchers and aspiring researchers must be always on trend to improve data extraction.

3.1.1.3 Amount of hashtag and Hashtag Variant

The amount of hashtag is unlikely influenced by the candidate’s popularity but is correlated to the candidate’s activity in the past, present, and his/her future activities. The issue on the amount of hashtags is that not every hashtag is created equal, there are several hashtag that has different variants such as #bbmsara, #bbmsara2022, #2022bbmsara. These hashtag variants must be considered in partner with trending hashtags in twitter.

3.1.2 Issues on TWINT API

The amount of tweets extracted by the TWINT is influenced by its age, it has been observed by the researchers that the amount of tweets extracted is lesser if a tweet’s age is more than two weeks. To be able to extract more data on a certain filter, being ahead of time and observing new trends in hashtags must be considered by the researchers to extract better data.

3.1.3 Translating to English

Mono language must be prioritized sentimental analysis, the researchers made use of Google Translate as the primary translating tool to English as Google Translate is defined as a multilingual neural machine translation service that takes context in translating a language to another. Automated translation to English is possible by the use of python and googletrans API.

3.1.3.1 Issues on Translating to English

3.1.3.1.1 The amount of informal texts on data gathered

Dialects that is impossible to translate to English, shortcut texts, grammatical errors, typographical errors, spacing, ‘taglish’, ‘konyo’ and others made the researchers to conclude that is must be impossible to translate a regular Filipino tweet to a good English sentence without supervising an algorithm.

3.1.3.1.2. Translating after data cleaning.

Translating the data after cleaning of noises must be a good idea. Upon manual analyzing of translated data, it has been observed by the researchers that certain words are translated to its literal English counterpart such as ‘lutang’ is to ‘float’. ‘lutang’ in context to present tweets refers to Leni’s ‘lutang’ moments during the debates and interviews and it is leaning towards negative sentiment but due to translation, the sentimental meaning of it has been lost.

3.2 Data Annotation

3.2.1 Manual Annotation of Data

Annotation of data sets involves classifying and labeling the tweets using three classes: positive, negative and neutral. This process is implemented to analyze the sentiments when training a model and to get acquire a better and accurate results.

In the study, the researchers used random sampling to select annotated 20% of tweets from each collected dataset. Understanding the nature of public sentiments, the researchers estimates at least 25% of data from the dataset are considered as bad data and will be discarded once the preprocessing has been done. Given this assumption, 25% of 20% is 5%, so an additional 5% is added for a total of 25% or 10,547 of data will be selected from the dataset to be used as the training and testing dataset. The tweets were manually annotated by the researchers to identify if they had positive, negative or neutral sentiment. This process is also significant in machine learning models to classify each sentiment and evaluate its performance accordingly.

Table 2. Total Annotated Data

CANDIDATES	Total Annotated Data	Sentiments count and percentage (%)		
		Positive	Negative	Neutral
LENI	6,878	5,449 (79.22%)	421 (6.12%)	1,008 (14.66%)
BONG BONG	2,779	919 (33.07%)	1,719 (61.86%)	142 (5.11%)
ISKO	890	682 (76.63%)	83 (9.33%)	125 (14.4%)
Total	10,547	7,050 (66.84%)	2,810 (21.07%)	1,275 (12.09%)

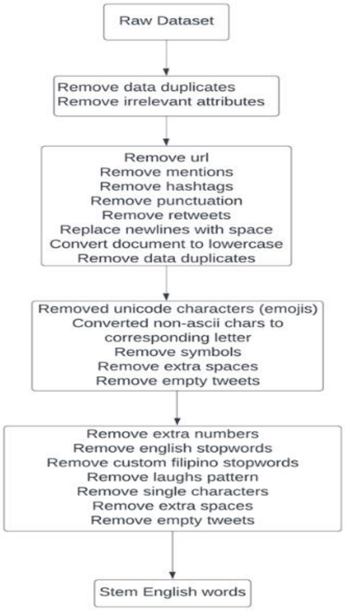
It can be observed that much of Leni dataset is skewed to positive sentiment with 79.22% coverage, similarly to Isko dataset with 76.63% coverage. Meanwhile, Bongbong dataset is highly skewed to the negative sentiment which is 61.86% of the whole dataset. This is a rare occurrence since public political tweet's nature are usually to show support and to spark an argument with other political parties, the Bongbong dataset that is highly skewed to negative sentiment shows that twitter users may have negative bias to the said presidential candidate.

The total positive sentiment covers 66.84% of the data, 21.07% in negative sentiment and a 12.09% for the neutral sentiment. Most of the negative sentiment comes from the Bongbong dataset which may affect the resulting machine learning algorithm's accuracy rating.

3.3 Data Preprocessing

The data gathered contained enormous amount of noise the sequence of data cleaning is shown in figure 2.

Figure 2. Data Cleaning and Stemming



The researchers utilized Microsoft Excel, base R and R libraries such as tm, tidytext, tidyverse, stringr for the data preprocessing. Initially, the researchers have compiled all the data cleaning in one code. Upon review of the dataset, there has been a lot of noise such as extra white spaces, single numbers, and single characters left from a data. The researchers concluded that removal of extra white spaces, numbers, characters, removal of data duplicates must be done for every stage to further clean the dataset.

3.3.1 Removal of irrelevant attributes and removal of data duplicates

The researchers took advantage of the Microsoft Excel's data duplicate removal and removing columns that is not needed by the sentimental classification for efficiency. Irrelevant attributes such as place, mentions, urls, photos, timezone, etc. and data duplicates that is resulted from spams from several users is also eliminated from the data set.

3.3.2 Noise Reduction

The decision to separate the noise reduction into two parts is because the researchers have observed that data duplicates frequently occurred after transforming texts.

3.3.2.1. Elimination of noises in a common tweet structure.

Elimination of URLs such as 'https://t.co/HgJDsNWtrl' have been eliminated using string patterns starting with 'https' or 'http'. Mentions such as '@bongbongmarcos' '@lenirobredo' have been eliminated using string patterns that starts in '@'. Any form of hashtag has been removed such as '#Lenlen' have been removed through the use of string pattern '#'+ string. Retweets are eliminated using string pattern 'RT', these are usually found at the beginning of a tweet.

The usage of emoji in a tweet is common for twitter users, it may be used in future studies in sentimental analysis but since this study is focused on texts, any emoji has been removed. Emojis are read as Unicode characters so a string pattern '<U+ >' have been used to eliminate such emojis.

Conversion of non-ascii characters to its alphabet counterpart. Example of this is 'â' to 'a'. It is uncommon for users to look at these types of characters to be stylish but some do, so it is considered to be converted for a cleaner data.

3.3.2.2. Elimination of noise characters

Noise characters such as punctuations, symbols, extra white spaces, new line characters conversion to space, single characters such as 't' and 'd' and numbers resulted from any form of text cleaning has been removed. Moreover, all the uppercase of the dataset have been converted to lowercase and empty tweets resulted from the data cleaning have been removed to further clean the dataset.

3.3.2.3. Elimination of Stop words

The researchers removed English stop words dictionary that is provided by the R library 'tm'. English stop words such as 'the', 'is', 'are' are removed from the document.

To remove Tagalog, stop words, the researchers manually checked the data set's texts using text frequencies. Tagalog stop words such as 'sa', 'akin', and 'ko' have been identified and removed from the data set as these words do not convey any sentiments.

3.3.2.4 Stemming

The researchers used R library 'tm' provides Porter's algorithm to stem a document, it is only limited to English words. English stemming is only used for the reason that there are only a few of public algorithms that focuses on outputting stemmed tagalog words. Upon the researcher's review on such work, these algorithms are incomplete, work in progress and thus using it for the study may further complicate the analysis. Another issue is the number of dialects used in the tweet, the use of Cebuano, Batangueno, Bicolano, and Bisaya is prevalent in the dataset, stemming such words may prove inaccurate and the potential of data sentiment may be lost. The researchers then decided to use the data as it is and manually set such words' polarity.

3.3.3 Preprocessing Results

After preprocessing, there is a total of 16.07% of data reduction dataset where it includes bad data quality such as spams, irrelevant tweets and etc. This is 9% lesser than the researcher's assumption of 25%. Table 3 shows the full summary of the preprocessing results which includes the data reduction in the positive, negative and neutral sentiments.

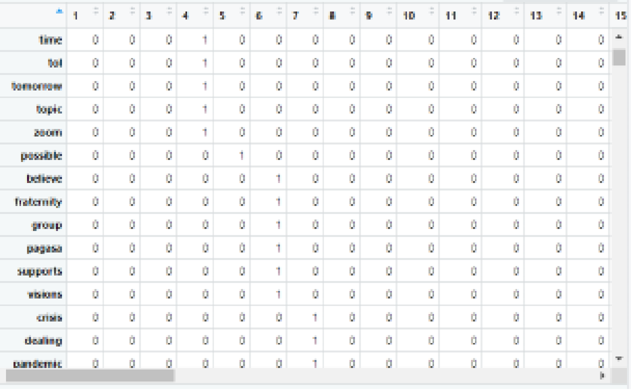
Table 3. Summary of Preprocessing Results

Annotated data	Data Count	Sentiments count and percentage (%)		
		Positive	Negative	Neutral
Before Preprocessing	10,547	7,050 (66.84%)	2,810 (21.07%)	1,275 (12.09%)
After Preprocessing	8,852	5,838 (65.95%)	1,911 (21.59%)	1,103 (12.46%)
Data Removed	1,695 (16.07 %)	1,212 (17.19%)	899 (31.99%)	172 (13.49%)

3.4 Data Representation

3.4.1 Term-document Matrix

Figure 3. Term-document Matrix



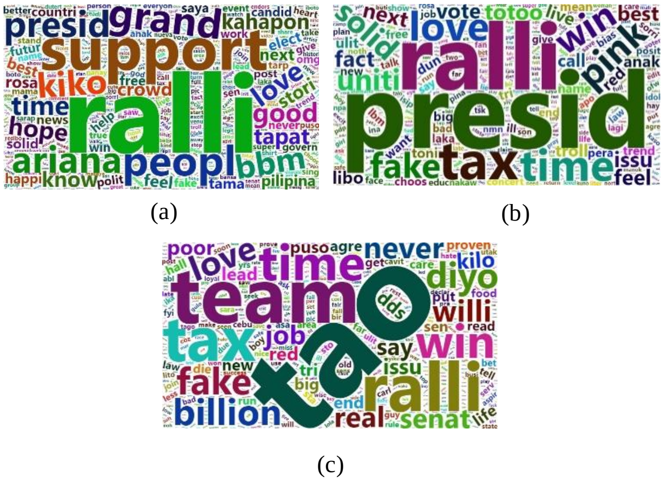
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
time	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
tol	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
tomorrow	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
topic	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
zoom	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
possible	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
believe	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
fraternity	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
group	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
pagasa	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
supports	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
visions	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
crisis	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
dealing	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
pandemic	0	0	0	0	0	0	1	0	0	0	0	0	0	0	

After preprocessing data, several methods are used in data mining to represent and analyze preprocessed data. One common example of displaying data is through a term-document matrix. A term-document matrix is a method of representing data in the form of a matrix. It is filled with weights that correspond to the term's importance in the specific document. A TDM also composed of rows and columns that identify and represent the tweets.

The researcher also used a function TermDocumentMatrix () from the text mining package to display the matrix for the top three presidential candidates in this year's election. The installation of package "tm" is significant in this method of running the function in the RStudio.

3.4.2 Word clouds

Preprocessed data analysis also includes the creation of a word cloud, also known as a text cloud. It is used in sentiment analysis as a virtual representation of text data. This stage also included the installation of packages such as the text mining package (tm) and the word cloud generator package (word cloud) to analyze texts and present the keywords as a word cloud. The bigger the words in the word cloud, the more frequently they occur in the corpus.



(a)

(b)

(c)

Figure 4. Wordcloud (a) "Leni" Robredo, (b) "Bongbong" Marcos and (c) "Isko" Moreno

The figure above shows the word cloud as a virtual representation of the words that appear more frequently for the three candidates. Using the packages "tm" and "word cloud", it can efficiently analyze and display the most frequent words for each candidate. Based on the figure above, the most frequent words for candidate are "ralli", "support", "grand", and "people", for candidate b, "presid", "ralli", "tax", and "time" and for candidate c, "tao", "team", "ralli", and "tax". The words "love" and "ralli" appeared frequently for all candidates. Different colors and font sizes were also used to classify each term based on its frequency.

3.5 Sentiment Classification

The manually annotated data will be used as the input for the machine learning classification techniques. The training and testing dataset with 8,852 instances will be used to evaluate the performance of the sentiment classification.

The researchers used four classification techniques for the sentiment classifications, the reason for this is that it can be applied applications for text classification. The researchers also used the repeated cross validation techniques as it has proven to increase the performance of each algorithm for this type of dataset. The split ratio for the training set will be two-thirds (2/3) of the total dataset and for the testing set, one-third (1/3) ratio will be used.

3.5.1 Naïve Bayes

Naive Bayes is a set of classification algorithms based on the Bayes Theorem. When it's used for text data analysis, the Naive Bayes classifier produces excellent results. The Naive Bayes algorithm calculates a probability based on the data set we provide. As a probabilistic classifier, the Nave Bayes classifier is used. It employs the concepts of mixture models to carry out the classifier. A mixture model can use the Bayes theorem to determine the probability of the component that it consists of in order to perform as a probabilistic classifier. A naive Bayes is also known as a simple Bayes or an independence Bayes. P stands for probability and is defined as follows:

$$P(y|z)=\frac{P(y|z)\,P(z)}{P(y)}$$

Above,
P (y | z) is the probability of class x. Where x is the target and predictory is the attribute.
P (z) is the prior probability of class.
P (y| z) is the probability of predictor of the given class.
P (y) is the prior probability of predictor.

3.5.2 Support Vector Machine

SVM is the most popular ML-based pattern classification technique available today. It was created in 1995 by Vapnik and is based on statistical learning theory. The primary goal of this technique is to use various types of kernel functions to project nonlinear separable samples onto another higher dimensional spaceThe score of the texts is also calculated, and the score is used as input to Support Vector Machine. However, text categorization may occasionally produce results. A text classifier comparison is required to determine which one is superior between texts. In this case, the performance measure is used [11].

3.5.3 Random Forest

Random forest is a type of automatic learning technique and was formally proposed by Leo Breiman and Adèle Cutler in 2001. The concepts of random subspaces and "bagging" are combined in this algorithm. The decision tree forest algorithm is trained using multiple decision trees that are fed slightly different subsets of data [12]. The outcome is determined by this algorithm based on the predictions of the decision trees. It predicts by averaging the output of various trees. The precision of the outcome improves as the number of trees increases.
The limitations of a decision tree algorithm are removed by using a random forest. It reduces dataset overfitting and improves precision.

3.5.4 Decision Tree Classifier

The Decision Tree algorithm is used to create classification and regression models. It is used to build data models that predict class labels or values for use in decision-making. The models are created using the system's training dataset (supervised learning). A decision tree was used to visualize decisions, making them easier to understand, and is thus a popular data mining technique.
The decision tree is also used as a classification algorithm after the feature vector has been created. The tweet is classified as positive, negative, or neutral by the decision tree. The proposed methodology's experimental results show significant success in terms of accuracy and sentiment analysis [13].

4. RESULTS AND EVALUATION

The total number of tweets collected for the three candidates is 42,196 using TWINT and a total of 8,852 are used for performance evaluation for the machine learning algorithms used in the study. The 8,852 tweets were already manually annotated and preprocessed to improve the performance of the algorithm applied in R. The results of applying each algorithm to the datasets were also demonstrated in R, allowing for comparison and analysis of the performance of the machine learning approaches used in the study.

Table 4 displays the different results of each classification model used in the study. Based on the results, the model based on Naïve Bayes has the highest accuracy rate of 100% among other classifiers. It was followed by the Decision Tree with overfitting with 68.42% and the random forest with 67.48%. Lastly, SVM has the lowest accuracy rate of 67.27% among the four algorithms used for the given dataset. The results show that Nave Bayes is the best performing approach for classifying sentiments on the collected data set, with 100% accuracy, precision, recall, and F1 score.

Table 4. Comparative Analysis of Results

Classifier Model	Naïve Bayes	SVM	Random Forest	Decision Tree
Accuracy	100%	67.27%	67.48%	68.42%
Precision	100%	96.925	96.66%	97.07%
Recall	100%	67.61%	67.85%	68.34%
F1 Score	100%	79.65%	79.73%	80.21%

The confusion matrix was also used to provide information about the performance of machine learning models on a test set with known labels. In other words, it provides insight not only about the errors made by a classifier model, but also about the types of errors made. The tables below show the confusion matrix of the four algorithms.

Table 5. Naïve Bayes Confusion Matrix

Naïve Bayes	Negative	Neutral	Positive
Negative	645	0	0
Neutral	0	360	0
Positive	0	0	1,919

Table 6. SVM Confusion Matrix

SVM	Negative	Neutral	Positive
Negative	107	7	59
Neutral	0	0	0
Positive	538	353	1,860

Table 7. Random Forest Confusion Matrix

Random Forest	Negative	Neutral	Positive
Negative	118	7	63
Neutral	1	0	1
Positive	526	353	1,855

Table 8. Decision Tree Confusion Matrix

Decision Tree	Negative	Positive
Negative	130	57
Positive	875	1,889

5. CONCLUSIONS AND FURTHER WORK

Text mining and sentimental analysis have enabled us to acquire, analyze, visualize and interpret data particularly during elections. Thus, this study aims to classify election-related public tweets into positive, negative and neutral and apply them to the machine learning algorithms to make comparison and evaluation of their performance. Manual annotation and preprocessing of data have also improved the ML algorithms which made it more efficient to classify the datasets. The result also shows that Naïve Bayes with a 100% accuracy rate was the most efficient and accurate ML approach among the other models.

For further work, it is better to have an in-depth investigation into other machine learning algorithms for this type of study. Since this paper describes a few examples of them, it is essential to study the various aspects of other ML approaches applicable and expand them further. Additionally, more research should be done to construct a larger number of datasets and improve data quality to get better performance for each ML algorithm.

6. ACKNOWLEDGEMENT

This work is produced for the subject Elective 2- Data Mining in Bicol University, College of Science, Computer Science and Information Technology Department.

7. REFERENCES

[1] DataReportal. 2022. Digital 2022: The Philippines. Retrieved March 28, 2022 from https://datareportal.com/reports/digital-2022-philippines

[2] Chauhan, P., Sharma, N., Sikka, G. 2021. The emergence of social media data and sentiment analysis in election prediction. Retrieved March 28, 2022 from https://link.springer.com/article/10.1007/s12652-020-02423-y

[3] A. Java, (2007, August). Why we twitter: understanding microblogging usage and communities, Retrieved March 28, 2022 from https://dl.acm.org/doi/abs/10.1145/1348549.1348556

[4] Cozma R., Chen K. 2011. Congressional Candidates' Use of Twitter During the 2010 Midterm Elections: A Wasted Opportunity?. In 61st Annual Conference of the International communication association,
Retrieved March 26, 2022 from https://www.semanticscholar.org/paper/Congressional-Candidates%E2%80%99-Use-of-Twitter-During-the-Cozma-Chen/e818d6cbefa29048ccd9afbd9504cd1ff5022f5f

[5] ProjectPro. 2021. Data Preprocessing - Techniques, Concepts and Steps to Master Website. Retrieved April 17, 2022 from, https://www.projectpro.io/article/data-preprocessing-techniques-and-steps/512

[6] Dařena, F., and Žižka J. 2015. Interdependence of text mining quality and the input data preprocessing. In Artificial Intelligence Perspectives and Applications Springer, Cham. Retrieved March 26, 2022 from https://www.springerprofessional.de/en/interdependence-of-text-mining-quality-and-the-input-data-prepro/2439698

[6] Le, B., Nguyen, H. 2018. Twitter Sentiment Analysis Using Machine Learning Techniques. In: Le Thi, H., Nguyen, N., Do, T. (eds) Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, vol 358. Springer, Cham.

[7] Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar and Mohammad Masum. 2019. Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. 8th International Conference on System Modeling and Advancement in Research Trends, 22nd–23rd November. Retrieved April 24, 2022 from https://www.researchgate.net/publication/342221481_Comparison_of_Naive_Bayes_and_SVM_Algorithm_based_on_Sentiment_Analysis_Usi ng_Review_Dataset

[8] Abdullah Alsaeedi, Mohammad Zubair Khan. 2019. A Study on Sentiment Analysis Techniques of Twitter Data, (IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 10, No. 2, 2019

[9] Yassine Al Amrani, Mohamed Lazaar, and Kamal Eddine El Kadiri. 2018. Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Procedia Computer Science Volume 127, 2018, Pages 511-520, Retrieved April 24, 2022 from https://www.sciencedirect.com/science/article/pii/S1877050918301625

[10] Agus Pamuji. 2021. Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment. Vol. 07, No.01, Februari 2021, Retrieved April 26, 2022 from, https://journal.uc.ac.id/index.php/JUISI/article/view/2084/1580

[11] Sandeep Kumar Satapathy, SatchidanandaDehuri, Alok KumarJagadev and Shruti Mishra. 2019. Chapter 1 – Introduction. Retrieved April 24, 2022 from, https://www.sciencedirect.com/science/article/pii/B9780128174265000016

[12] Yassine Al Amrani, Mohamed Lazaar, Kamal Eddine El Kadiri. 2018. Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis, Procedia Computer Science,Volume 127,2018, Pages 511-520, Retrieved April 26, 2022 from,https://www.sciencedirect.com/science/article/pii/S1877050918301625

[13] R. Bibi, U. Qamar, M. Ansar and A. Shaheen, Sentiment. 2019 Analysis for Urdu News Tweets Using Decision Tree, 2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA), 2019, pp. 66-70, Retrieved April 26, 2022 from https://ieeexplore.ieee.org/document/8886788