A Textual Analysis of the 2022 Philippine Presidential Elections

A THESIS PRESENTED TO THE

DEPARTMENT OF INFORMATION SYSTEMS

AND COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

BY

MALLORY BELICIA T. MAROKET

KRISTINE KAITLYN MARIANNE W. MERCADO

ANDREA FRANCESCA M. PILAPIL

QUEZON CITY, PHILIPPINES

2021-2023

ABSTRACT

The advancements in technology and the use of social media have exponentially grown in the 21st century. Although, with these developments comes its weaponization against online users. One of the greatest threats when it comes to social media use is the exposure to and spread of misinformation. This study aims to make use of machine learning to classify whether political tweets about the 2022 Philippine Presidential Elections contain factual information or misinformation and to identify the characteristics of political misinformative tweets through exploratory data analysis. The study uses various machine learning classification algorithms trained on TF-IDF and Word2Vec vectorizations of the cleaned political tweets. To train the model, the researchers classified the tweets into three categories: factual (1), misinformative (0), and N/A. Using a 76-24 split to training and testing data, the best performing model was the TF-IDF Support Vector Machine model with a mean macro F1 score of 66% (+/- 3%) and mean accuracy of 70% (+/- 3%). From the classified tweets, these were segmented into the presidential candidate being talked about where the sentiment and topic of the tweet were also obtained through an exploratory data analysis, to infer their overall impact during the campaigning period.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The advancements in technology and the use of social media have exponentially grown in the 21st century. Although, with these developments comes its weaponization against online users. One of the greatest threats when it comes to social media use is the exposure to and spread of misinformation. In a previous study [4], there has been evidence of organized social media manipulation campaigns in up to 70 countries.

Many studies have been conducted in the field of online misinformation dissemination though few have focused on its occurrences and effects during national election periods, even fewer in the Philippine context. A study conducted by Kajimoto et al. (2019) has found that the Philippines is among the countries that employ trolls to amplify disinformation messages and bully critics or dissenters [10]. In this regard, the need to study misinformation dissemination in the Philippine context becomes even more crucial.

Many studies have been conducted in the field of online misinformation dissemination, though few have focused on its occurrences and effects during national election periods, even fewer in the Philippine context. Another study has found that the Philippines is among the countries that employ trolls to amplify disinformation messages and bully critics or dissenters [10]. In this regard, the need to study bot behavior and misinformation dissemination in the Philippine context becomes even more crucial.

Additionally, numerous studies have also claimed to have difficulty in collecting adequate amounts of data from social media platforms. Methods in data collection from similar studies have been found by researchers to be time-consuming and challenging. In a similar study, one of the challenges the researchers encountered had to do with Twitter's sampling algorithm not being uniformed and the periods of data collection being uneven which may have led to disparities in dataset sizes [30]. Another study manually collected 100 news items from online sources which were further reduced to only 60 news items in total [8]. The researchers also recommended expanding the focus of the study by examining other misinformation contexts and having a larger-scale dataset.

With these gaps in mind, this study aims to fill these by understanding the interactions between social media users more specifically in the Philippine context. Additionally, the researchers also hope to refine the methodology of previous studies in identifying misinformation and collecting data. The study will strive to do so by focusing on one specific period in time for consistency in data. This period in time is the 2022 Philippines national elections, from the beginning of the official campaign period in February 2022 until its last day in

May 2022.

1.1 Research Questions

Due to the Philippines enforcing indoor directives in an attempt to curb COVID-19 transmissions in the country, this has brought people to rely on social media for any information, such as the news. It is important to be cautious of the veracity of the information and credibility of its sources since these cannot be immediately identified by an internet user. With this, the researchers aim to answer the following main question and sub-questions through this study:

How can the characteristics of misinformation on Twitter surrounding the 2022 Philippine Presidential Elections be determined?

1. How can political tweets be classified into factual information and misinformation?
2. What are the characteristics of political tweets that contain misinformation?
3. What are the characteristics of the political tweets related to each presidential candidate?

1.2 Objectives of the Study

By the end of the study, the researchers aim to achieve the following goals:

1. To create machine learning models that classify whether a political tweet contains factual information or misinformation

2. To characterize political factual and misinformative tweets related to each presidential candidate

1.3 Scope and Limitations

The study focuses on the spread of misinformation during the 2022 Philippine presidential elections. Given this, the data collected will be during the campaigning period from February 8, 2022, to May 7, 2022, from public accounts on Twitter that have used the query #Halalan2022 and mentioned the names of the top 5 presidential candidates.

Moreover, misinformation includes fake news, rumors, deception, clickbait, satire, and manipulated content. The intention behind dissemination does not matter; Whether shared deliberately (disinformation and malinformation) or inadvertently (misinformation), the term "misinformation" in the study pertains to any kind of wrong and false information.

The machine learning models used to classify factual information and misinformation are Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbors, and Decision Tree.

1.4 Significance of the Study

In the last decade, we have witnessed an unprecedented rise in social media platforms and large-scale efforts to weaponize them in order to manipulate public opinion. Internet users have become more susceptible to misinformation. Through this research, internet users can have an increased public understanding of how social media platforms can be weaponized in order to manipulate public opinion. Our methodology for misinformation detection may also serve as a tool for those who wish to detect misinformation on Twitter. Additionally, this research may shed light on the current political context of our country.

Electoral misinformation represents a major threat to democracies worldwide as it is an intentional way of distorting the truth, in order to promote political candidates or influence political ideologies. Thus, this study would be of immense benefit to voters in particular and the general public at large, who seek authentic, reliable and verifiable information on political candidates. This research shall increase the awareness of voters of how misinformation online can affect electoral discourse, as well as help them make informed decisions in future elections.

Moreover, this study shall contribute to the literature related to the dissemination of online political misinformation. This research provides insights for other researchers who wish to conduct further studies on the related field.

CHAPTER II

REVIEW OF RELATED LITERATURE

Misinformation is a growing problem in today's world and can have serious consequences. It refers to false or inaccurate information that is spread intentionally or unintentionally. Misinformation can lead to negative effects on society, including social unrest, political polarization, and even health risks. Therefore, detecting and combating misinformation has become a crucial issue.

In recent years, word embedding techniques such as TF-IDF and Word2vec have emerged as powerful tools for natural language processing and text classification. These techniques help to capture the meaning of words in a more sophisticated way than traditional methods, allowing for a more accurate classification of texts.

Moreover, machine learning classification algorithms, such as Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree, and K-Nearest Neighbors have also been widely used in the classification of misinformation. These algorithms can help to identify patterns in text data and accurately classify them into different categories.

Exploratory data analysis is also an important step in the methodology of detecting and combating misinformation. By visualizing and analyzing the data, researchers can gain insights into the characteristics of misinformation and develop more effective approaches for its detection and prevention.

This chapter will provide an overview of the current methods for detecting and combating misinformation, with a particular focus on the use of word embedding techniques and machine learning classification algorithms.

2.1 Misinformation

2.1.1 Definition of Misinformation

The rise of social media platforms revolutionized how we create, share, and consume information, greatly improving its transmission velocity and available volume. However, while social media platforms have helped in creating a network for the widespread consumption of news and information around the globe, its unregulated and decentralized environment allows the mass proliferation of misinformation online[5]. Before delving into the discussion about online misinformation, it's important to clarify other similar terms that have been used interchangeably with misinformation such as rumor, fake news, false information, deception, hoaxes, spam, opinion, and disinformation. Despite being similar, there exist notable differences among them in terms of the degrees of wrongness, the contexts of usage, and the functions of serving for different propagation purposes[28].

The term Rumor pertains to a story of circulating information from person to person whose veracity status is unreliable and often unverified[28]. Fake news is differentiated by the content that mimics news media which intentionally misleads the readers, and it is verifiably false[28]. With the growing use of the term “fake news” in popular culture, academics have raised concerns about this umbrella term because of its conceptual ambiguity and misuse by political actors. In lieu, the terms disinformation and misinformation have been preferred in academic discussions[13]. Deception is generally defined as an intentionally misleading statement, as a means of conceptualizing deceptive communication both implicitly and explicitly[28].

Although both disinformation and misinformation refer to factually incorrect information, they are primarily distinguished by intentionality. Disinformation refers to deliberate attempts to manipulate public opinion through the use of false information, while misinformation pertains to the unintentional spread of false information[13]. Moreover, misinformation spreads faster, deeper, and broader on social media than factual information[5]. Due to the high volume of information that we consume when we use social media, humans have a limited ability to distinguish true information from misinformation[31]. In this research, we used misinformation because this is the most neutral and value-free term to describe inaccurate information in the context of communication research. Although misinformation is defined to be unintentional, it is difficult to detect the intention of information creators due to the insufficient information of such metadata. Therefore, in the context of this study, we define misinformation as false online information without the intentionality aspect.

2.1.2 Effects of Misinformation

Online Social Networks (OSNs) have evolved into major means of information dissemination. Analysis from Kepios shows that there are 4.65 billion social media users around the world in April 2022, equating to 58.7 percent of the total global population[17]. However, one major challenge is that the information communicated through the network is not always credible. The rise of misinformation often circulated in various social media platforms has raised growing concerns not only among policymakers and civil society groups, but also among citizens[13]. The continuous rise of misinformation has led researchers to argue that we are living in a "misinformation society"[13]. This claim is also reflected in a Reuters survey, which shows that citizens globally are concerned about the growing incidences of misinformation[12]. Twitter has evolved into one of the most popular microblogging services. Twitter offers users the opportunity to report a tweet as spam, compromised, or abusive. However, tweets containing misinformation may still not be identified.

The widespread misinformation causes a major social problem, as it damages the public trust in factual information, harming the democracy, justice, economy, public health, and security[5]. The spread of misinformation, rumors and fake news has impacted some influential and important events like elections and international disputes. The adversarial use of social media to spread deceptive or misleading information poses a political threat. For instance, during the 2016 US presidential election, as many as 529 different rumors were spreading on Twitter, and approximately 19 million malicious bot accounts published or retweeted tweets supporting Trump or Clinton, which potentially influenced the election.[9]. One study found that online misinformation was linked to lower trust in mainstream media across party lines[16]. However, for moderates and conservatives, exposure to fake news predicted a higher confidence in political institutions. Online misinformation accelerates propaganda, creates anxiety, induces fear, and sways public opinion; thereby having adverse societal impacts[6]

2.2 Word Embedding Techniques

Word embedding techniques are used to represent words mathematically. It does so by tokenizing each word in a sequence (or sentence) and converting them into a vector space[1]. Word embeddings aim to capture the semantic meaning of words in a sequence of text. There are many techniques available at our disposal to achieve this transformation. In this study, we will be covering two: TF-IDF and Word2vec.

2.2.1 Term Frequency-Inverse Document Frequency

TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents (also known as a corpus) [21]. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. In python, TF-IDF values can be computed in the library scikit-learn, which contains a module called TfidfVectorizer [20].

TF-IDF for a word in a document is calculated by multiplying two different metrics, TF (Term Frequency) and IDF (Inverse Document Frequency) [19]. The Term Frequency score is based on the frequency of a particular term you are concerned with relative to the document. Words are counted for their number of occurrences in the documents. While the Inverse Document Frequency is based on how common (or uncommon) a word is amongst the corpus. IDF score calculates the rarity of the words. Moreover, IDF captures words that are rarely used in the corpus that may hold significant information.

Multiplying these two numbers results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document. To put it in more formal mathematical terms, the TF-IDF score for the word t in the document d from the document set D is calculated as follows [27]:

$$tf\,idf\,(t, d, D) = tf\,(t, d)\;.\;idf\,(t, D)$$

Figure 2.1: TF-IDF Formula [27]

While TF and IDF are computed as follows:

$$tf\,(t, d) = log\,(1 + freq\,(t, d)) \qquad idf\,(t, D) = log\,\left(\frac{N}{count\,(d \in D{:}t \in d)}\right)$$

Figure 2.2: TF and IDF Formulas [27]

While machine learning algorithms traditionally work better with numbers, TF-IDF algorithms help them decipher words by allocating them a numerical value or vector [27]. This has been useful for machine learning, especially in fields related to NLP such as text analysis. In text analysis with machine learning, TF-IDF algorithms help sort data into categories, as well as extract keywords.

2.2.2 Word2vec

Word2vec is one of the most popular implementations of word embedding. Word2vec is a two-layer neural net that is used to create a distributed representation of words into numerical vectors [11]. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus. Word2vec converts text into vectors that capture semantics and relationships among words.

Word2vec is not a singular algorithm, rather, it is a family of model architectures and optimizations that can be used to learn word embeddings from large datasets [29]. Word2vec utilizes two architectures, the Continuous Bag-of-Words model (CBOW) and the Skip-gram model [29]. The continuous bag-of-words model predicts the middle word based on surrounding context words. The context consists of a few words before and after the current word. This architecture is called a bag-of-words model as the order of words in the context is not important. While the Skip-gram model predicts words within a certain range before and after the current word in the same sentence.

The purpose of Word2vec is to group the vectors of similar words together in vector space. That is, it detects similarities mathematically [14]. Given enough data, usage, and contexts, Word2vec can make highly accurate guesses about a word's meaning based on past appearances. Those guesses can be used to establish a word's association with other words, or cluster documents and classify them by topic.

2.3 Machine Learning Classification Algorithms

2.3.1 Logistic Regression

Logistic regression is a binary classification method through finding the best S-shaped logistic function between two labels. To determine the label of a sample, its continuous and discrete data can be used to train the model, which "makes it a popular machine learning method" [23]. Additionally, logistic regression can also be used to identify which types of data is useful or not useful in classifying samples.

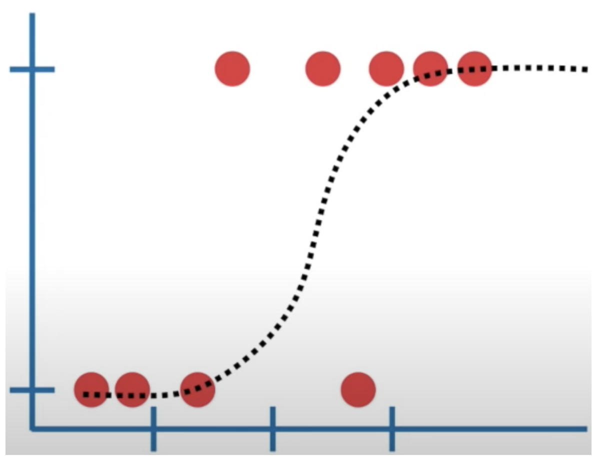

Figure 2.3: Sample Logistic Regression [23]

2.3.2 Naïve Bayes

Naïve Bayes is a classification method that makes use of probabilities based on the number of appearances of a token to determine the label. However, this method is called naïve because it does not value word structure and treats language "like it is just a bag full of words" [26], which is not always the case for all text classifying situations.

2.3.3 Support Vector Machine

Support Vector Machine (SVM) is a binary classification method through finding the best fit threshold between two labels on a cartesian plane. The threshold may be a line for two-dimensional data, a plane for three-dimensional data, and a hyperplane for higher dimensions of data. In building the machine, it "starts with data in a relatively low dimension, then moves into a higher dimension" [25] where the threshold is placed between two labels.

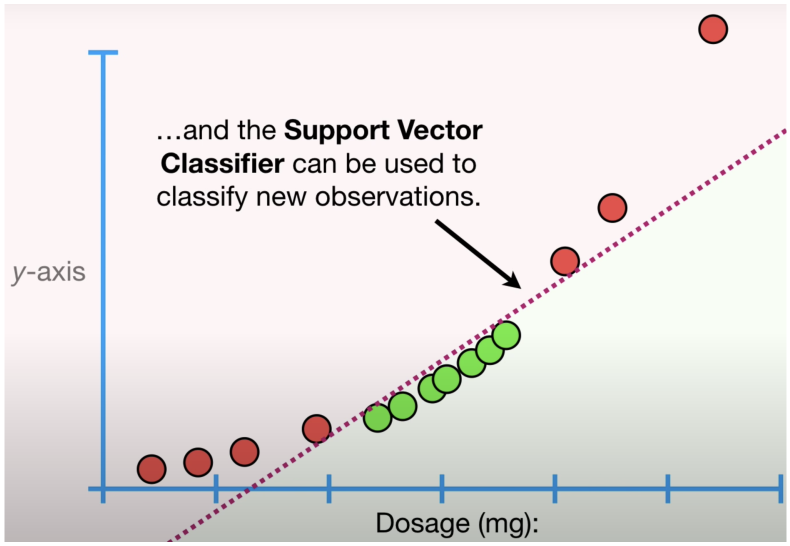

Figure 2.4: Sample Support Vector Machine [25]

Numerous previous studies that have compared different machine learning classification methods on identifying misinformation have arrived at the result that SVM has performed the best.

One of the studies [2] used 15,952 misinformation-related tweets from the 2013 Boston Marathon bombing, 2017 Manchester Arena bombing, 2017 Hurricane Harvey, 2017 Hurricane Irma, and 2018 Hawaii ballistic missile false alert. They tested and compared various machine learning classification methods including K-nearest neighbors, decision tree, random forest, SVM, and MLP. Of all the tests, the best performance came from SVM, with a macro-average F1 score of 0.872.

Seeing this result, they continued to test the model through lowering the training data to see how this would affect its performance. Even with a less significant amount of training data at 0.10, SVM still achieved a macro-average F1 score of 0.787.

Afterwards, they tested the SVM model on the six individual crisis events and compared these with their implementation of the Naïve Bayes, another machine learning classification method. The results showed that the two had a wide gap in its performance, the macro-average F1 scores of SVM performing 0.285-0.468 better than that of Naïve Bayes.

Another study [3] made use of Naïve Bayes, MLP, and SVM. The confusion matrix showed that all three of the models performed well at around 0.98, but the highest F1 score was that of SVM with 0.9896. The researchers also put importance on the fact that making use of machine learning algorithms works well only when the training data is labeled correctly and on the necessity to continuously retrain the model with contemporary data.

2.3.4 K-Nearest Neighbors

K-nearest neighbors is a classification method that starts with a dataset that has already been classified into categories, visualized on a scatter plot based on their characteristics. The sample is then placed on the scatter plot and depending on the value of K, the K number of plots nearest the sample is identified. The label of the most number of neighboring plots to the sample is then the label of the sample itself.

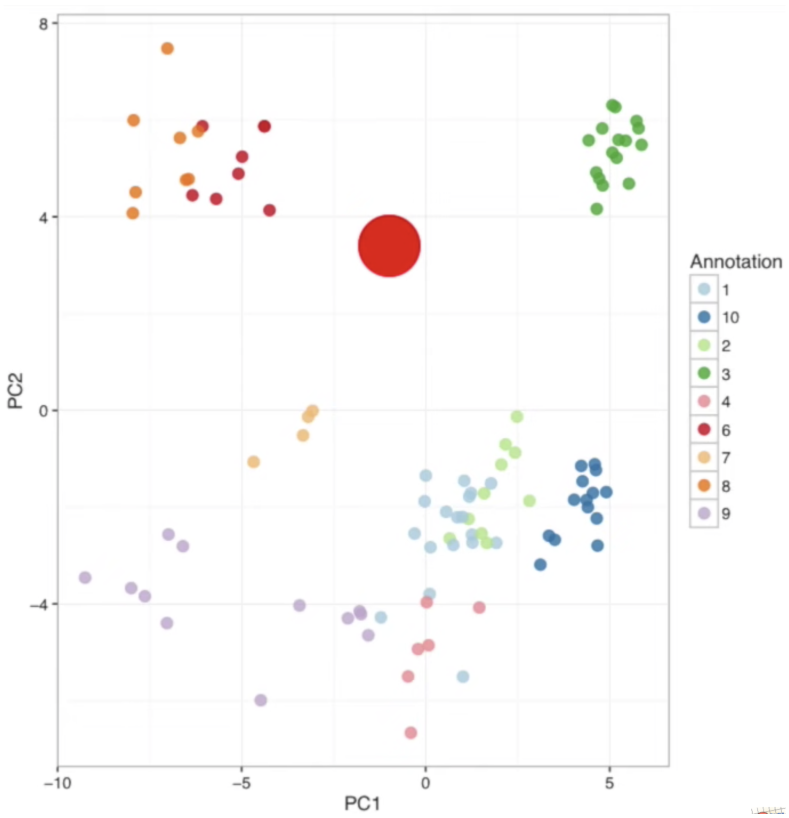

Figure 2.5: Sample K-Nearest Neighbors [22]

2.3.5 Decision Tree

Decision tree is a classification method through determining whether a statement is true or false which will then lead to either another statement or a classification based on the answer. The statements here are the data or characteristics of the sample. A decision tree starts with a root node containing the initial statement, and under the root node are called internal nodes or the branches containing another statement, or the leaf nodes or the leaves containing the final classification. However, decision trees "work great with the data used to create them, but they are not flexible when it comes to classifying new samples" [24].

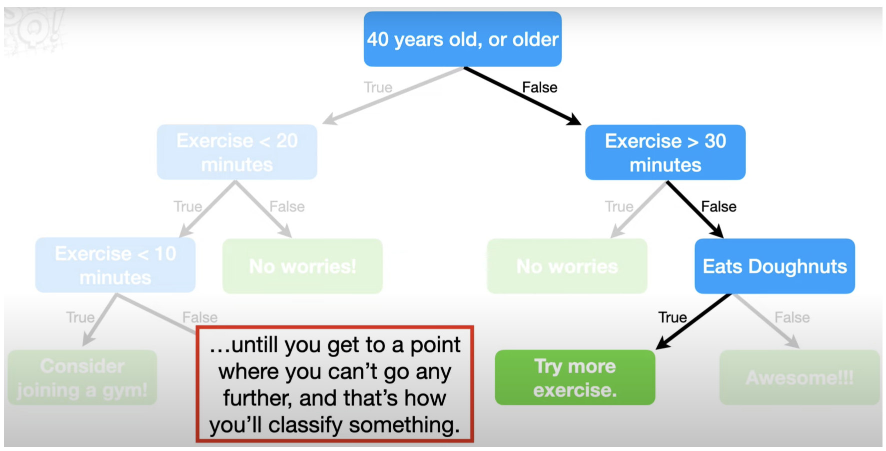

Figure 2.6: Sample Decision Tree [24]

2.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations [18]. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them.

It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, and find interesting relations among the variables [7]. It is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions [15].

CHAPTER III

METHODOLOGY

The methodology of this study will include four major phases: (1) Data Collection, (2) Data Cleaning, (3) Manual Classification, (4) Vectorization, (3) Machine Learning Models, and (4) Exploratory Data Analysis. Refer to Figure 3.1 for the flowchart of the methodology.

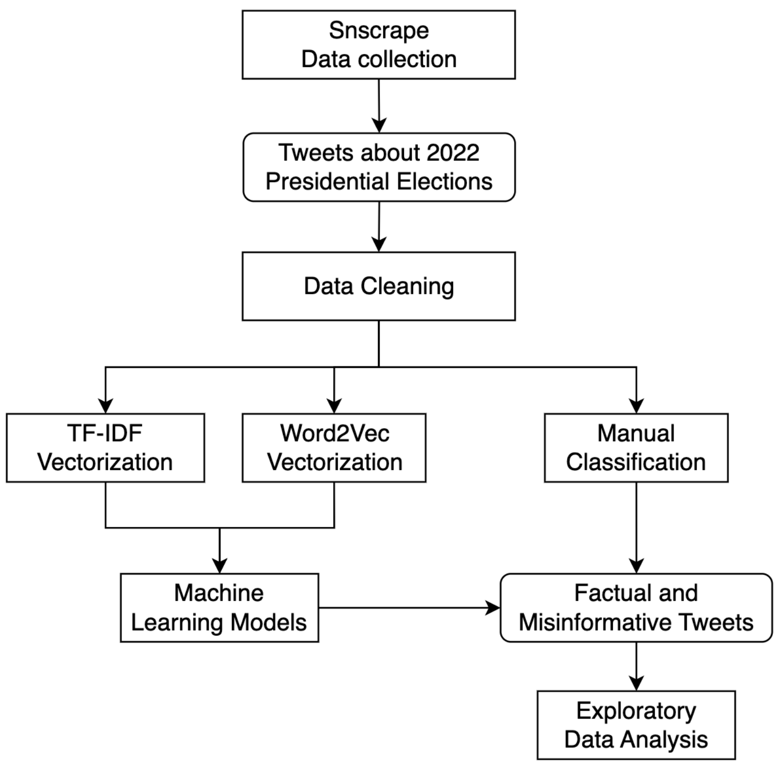

Figure 3.1: Methodology Flowchart

Additionally, a framework for these phases and their relation to the research questions are illustrated in Table 3.1.

Research Question	**Input**	**Process**	**Output**
How can political tweets be classified into factual information and misinformation?	Political tweets	Machine learning classification algorithms	Categorization of factual and misinformative tweets
What are the characteristics of political tweets that contain misinformation?	Misinformative political tweets	Topic clustering	Relevant topics, Tokens that contribute to the misinformative classification
What are the characteristics of the political tweets related to each presidential candidate?	Political tweets	Exploratory data analysis (EDA), Topic clustering	Characteristics of tweets per candidate

Table 3.1: Methodology Framework

3.1 Data Collection

The researchers will make use of Snscrape to perform tweet extraction, including its user-related and tweet-related metadata. To filter the tweets, they shall put a specific date range and location, which will be the campaigning and election period from February 2022 to May 2022 in the Philippines.

A summary of the restrictions and a list of sample keywords to filter election-related tweets are presented in Table 3.2.

Location	Philippines
Date range	February 8, 2022 to May 7, 2022
Query	#Halalan2022
Keywords	leni, robredo, bongbong, marcos, bbm, ping, lacson, manny, pacquiao, pacman, isko, moreno, or yorme

Table 3.2: Tweet Extraction Filters

3.2 Data Cleaning

To prepare the tweets to be fed into the machine learning classification algorithms, the researchers will implement text cleaning techniques, such as converting all text to lowercase, and eliminating non-alphanumeric characters, hashtags, duplicates, links, and white spaces. This is done to ensure that the extracted data is free from extraneous noise and irrelevant information, and that the data is consistent and reliable, improving the accuracy and efficiency of the machine learning model. The cleaned tweets shall then be used for the vectorization processes in getting the parameters of the model. An instance of the uncleaned and cleaned tweet is presented in Table 3.3.

Uncleaned tweet	**Cleaned tweet**
"Layout and display name changed temporarily to show support for Vice President Leni Robredo as she will run as the next President of the Philippines. #LetLeniLead #Halalan2022 https://t.co/UWYnBUP0vB"	layout display name change temporarily show support vice president leni robredo run next president philippines

Table 3.3: Tweet Cleaning Example

3.3 Manual Classification

Labeling misinformation is challenging, more so because misinformation is not easy to specify and has to be proven. In order to systemize this, a complete tagging manual containing specific annotation guidelines has been used as a basis to properly label the tweets. The annotators, or the researchers, will analyze each tweet with the help of URLs given and will rate each tweet as misinformative (0), factual (1), and N/A accordingly. The annotation guidelines the researchers have provided are as follows:

3.3.1 Annotation Guidelines

The annotation guidelines used to manually classify the tweets are described below:

A URLs list of verified Philippine news media like CNN Philippines, Rappler, ABS CBN News, GMA News, The Philippine Star, Manila Bulletin, the Manila Times, Philippine Daily Inquirer, Business Mirror, BusinessWorld, and other Philippine fact-checking websites, namely Vera Files, PressOnePH, and TsekPH were considered. Rating of each tweet has been done based on the following question:

Question the Source: Label the tweet based on what the tweet is trying to say or claim, and how factual its claim is. Choose one of the below labels for the tweet:

- True: if the primary elements of the claim are verifiably true.
  - Mostly true: if the primary elements of a claim are demonstrably true,

but some of the ancillary details surrounding the claim may be inaccurate.

  - Debunk: Tweet calls out or debunks inaccurate information.

- False: if the primary elements of a claim are false or conspiratorial.
  - Fake news, rumors, deception, clickbait, satire, and manipulated content
  - The intention behind dissemination does not matter (misinformation, disinformation, malinformation)
  - Mostly false: if the primary elements of a claim are false, but ancillary details may be accurate.

- N/A: if it is not one of the above, for example:
  - Unproven: if it has insufficient evidence to make a judgment
    * Ex. The tweet is referring to a deleted image or video
    * Ex. it uses a language (i.e., Bisaya) or references that are difficult to understand
  - Opinions
    * Ex. Merely asserting that they will/will not vote for a certain candidate or listing why they are/are not voting for them

- Tweets about previous elections or non-electoral candidates
- Tweets that are not understandable by English or Tagalog

Annotation steps:

1. Read the tweet text.
2. Determine if the tweet should be labeled as True, False, or N/A based on its source or other extra information provided in the tweets.
3. After discerning the proper label for the tweet, follow the legend below to properly tag the tweet and input this in the appropriate Google Sheet document:

Label	**Classification**
True	1
False	0
N/A	N/A

Table 3.4: Tagging Legend

3.4 Vectorization

To render the tweets machine-readable, the researchers will perform vectorization of the tweets using the TF-IDF and Word2vec vectorization methods.

3.4.1 TF-IDF

For the TF-IDF vectorization, the SciKit Learn library was used to import the built-in TF-IDF vectorizer for the vectorization of the dataset. A function was created using the vectorizer function to obtain the word vector score and token feature of each tweet in the dataset. The vectorization process involves computing the TF-IDF score for each word in the tweet based on its frequency in the tweet and its inverse frequency in the corpus of tweets.

3.4.2 Word2vec

For the Word2Vec vectorization, the FastText library was used to import a pre-trained Tagalog model. FastText generates word embeddings, which are dense vector representations of words. These embeddings capture the semantic and syntactic information of the words, and they can be used as input features for the machine learning models. The Tagalog model was loaded and applied to the collected dataset in order to obtain the sentence vector score of each tweet.

3.5 Misinformation Detection

The researchers will utilize the following supervised machine learning algorithms in order to classify the tweets into misinformation or factual information: Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), K-Nearest

Neighbors, and Decision Tree. This is to test out which one would perform best alongside the vectorization methods.

The numeric values from each vectorization method will be used to train their corresponding machine learning models. The TF-IDF dataset and FastText dataset will both be split into 76% training and 24% testing datasets. A validation dataset will not be implemented for this study because the values that will be used to predict the classification are not hyperparameters but only parameters, produced by the vectorization method and thus cannot be further reconfigured.

To evaluate the performance of the models, metric scores such as accuracy and macro F1 will be collected. Additionally, cross-validation shall be performed to find the mean of those metric scores and their standard deviation.

3.6 Exploratory Data Analysis

The exploratory data analysis was conducted using Python notebooks and several natural language processing libraries such as the Natural Language Toolkit (NLTK), spaCy, TextaCy, vaderSentiment, and pyLDAVis.

An NLP pipeline was created to identify the verbs, adjectives, and sentiments of the entire corpus. The dataset was also segmented according to each presidential candidate in order to identify the specific terms and sentiments for each candidate.

Finally, for the topic modelling, the process was finetuned to identify three main topics and the Top 30 most salient terms of each using the pyLDAvis library.

CHAPTER IV

RESULTS AND DISCUSSION

In this section, the summary of the results and data obtained from this study's methodology will be interpreted and discussed with the use of graphical models.

4.1 Manual Classification

Using the tagging manual as a reference, the researchers successfully classified a total of 11356 tweets pertaining to the 2022 Philippine presidential elections. Subsequently, duplicated tweets were eliminated, downsizing the corpus to 10350 tweets. The resulting dataset contained 4,995 factual tweets, 1,281 misinformative tweets, and 4,074 tweets in the N/A category, which was used to train and evaluate the machine learning models.

Classification	**Total Count with Percentage**
Factual (True/1)	4995 (48.26%)
Misinformative (False/0)	1281 (12.38%)
N/A	4074 (39.36%)

Table 4.1: Manual Classification Summary

4.2 Vectorization

4.2.1 TF-IDF

The TF-IDF vectorization process resulted in a 10350 × 3000 matrix representing the term frequency by inverse document frequency for the maximum number of features in the cleaned tweets. Without specifying the max_features parameter, there would be an excessive number of 18074 features. Because of this, the researchers chose a value of 3000 to avoid overfitting but at the same time have a wide selection of words to better accurately classify new tweets not in the original dataset.

If the term is common and appears in several tweets, the resulting value of the term will approach 0, otherwise, it will approach 1. The higher the score, the more relevant the term is in relation to the entire corpus. Figure 4.1 shows the partial results of the TF-IDF matrix.

	aatras	aayusin	abalos	abangan	abella	able	abogado	abortion	about	abra	...	youth	youtube	yrs	yun	yung	zamboanga	zero	zoom	zubiaga	zubiri
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10345	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10346	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10347	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10348	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10349	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

10350 rows × 3000 columns

Figure 4.1: TF-IDF Results

4.2.2 Word2vec

For the Word2vec vectorization process, the pre-trained Filipino FastText model was loaded and vectorized the tweets, resulting in a 2644 by 300 matrix representing the vectors of the sentence from the cleaned tweets. Figure 4.2 shows the partial results of the FastText matrix.

	0	1	2	3	4	5	6	7	8	9	...	290	291	292	293	294	295
0	0.012728	0.011190	0.032246	-0.017182	-0.054768	-0.016959	-0.004939	0.014671	0.002839	0.024351	...	-0.009520	0.022536	0.005348	-0.018480	0.025240	-0.042913
1	0.006445	0.017245	-0.007372	-0.005195	-0.046712	-0.000827	0.005396	-0.002076	0.025854	0.004707	...	-0.014660	0.014361	0.001755	0.042996	0.011142	-0.023046
2	-0.007537	0.020222	0.006260	0.000311	-0.051977	0.007121	0.012173	0.026976	0.000684	0.001145	...	-0.035879	0.000298	-0.014085	0.010937	0.010970	-0.019324
3	0.012362	-0.019312	0.006466	0.018907	-0.097392	-0.019668	-0.022090	-0.016476	-0.014075	-0.024208	...	0.066857	-0.013911	0.014908	0.007774	0.032452	-0.025076
4	-0.003331	0.034356	-0.025789	0.022169	-0.057575	-0.027852	0.007822	0.002140	-0.003544	0.029548	...	-0.037165	-0.001748	0.016419	0.029753	0.040884	0.004189
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10345	-0.009527	0.002800	-0.006713	-0.007107	-0.084784	-0.020310	-0.002018	-0.025998	0.025763	-0.022773	...	0.024675	-0.011581	-0.018253	-0.018903	0.053307	-0.016374
10346	-0.006032	0.030676	-0.028189	-0.036794	-0.043573	-0.021963	0.012410	0.007173	0.027786	0.001294	...	-0.025239	0.008455	0.025012	0.062420	0.036645	0.012882
10347	0.006002	0.037119	0.015261	0.000547	-0.069316	0.022650	0.036634	-0.035786	-0.029726	0.007165	...	-0.032007	0.002240	0.010663	0.019667	0.058877	-0.015726
10348	0.008405	0.023827	0.011798	-0.005126	-0.077323	-0.019237	-0.008696	0.033981	-0.005272	-0.002755	...	-0.023424	0.022146	-0.011804	0.023771	0.028005	-0.052496
10349	-0.000583	0.031025	-0.008970	-0.011674	-0.049267	0.003925	0.021657	0.011299	0.007033	0.007686	...	-0.039332	-0.004006	0.030815	0.040942	0.028770	-0.010690

10350 rows × 300 columns

Figure 4.2: Word2vec Results

4.3 Misinformation Detection

For each of the vectorizers, namely TF-IDF and Word2vec or FastText, five machine learning models were created: Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Decision Tree. The corpus was partitioned into a 76% training set and 24% testing set. To evaluate the models, confusion matrices were first plotted, where the number of correct and incorrect classifications could be obtained.

4.3.1 Support Vector Machine Model

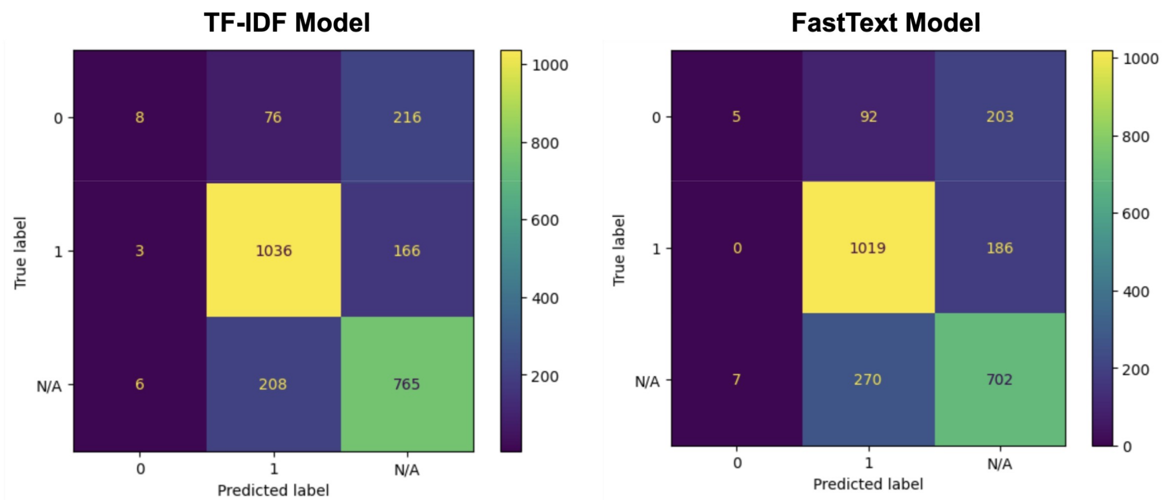

Figure 4.3: Support Vector Machine (SVM) Confusion Matrices

	TF-IDF Model	**FastText Model**
True Positives	1036 (41.71%)	1019 (41.02%)
True Negatives	8 (0.32%)	5 (0.2%)
True N/A	765 (30.8%)	702 (28.26%)
False Positives	284 (11.43%)	362 (14.57%)
False Negatives	9 (0.36%)	7 (0.28%)
False N/A	382 (15.38%)	389 (15.66%)

Table 4.2: Support Vector Machine (SVM) Results

As seen in Figure 4.3, the TF-IDF model was able to predict 1036 True Positives, 8 True Negatives, 765 True N/A, 284 False Positives, 9 False Negatives, and

382 False N/A. On the other hand, the Word2vec or FastText model was able to predict 1019 True Positives, 5 True Negatives, 702 True N/A, 362 False Positives, 7 False Negatives, and 389 False N/A.

4.3.2 Logistic Regression Model

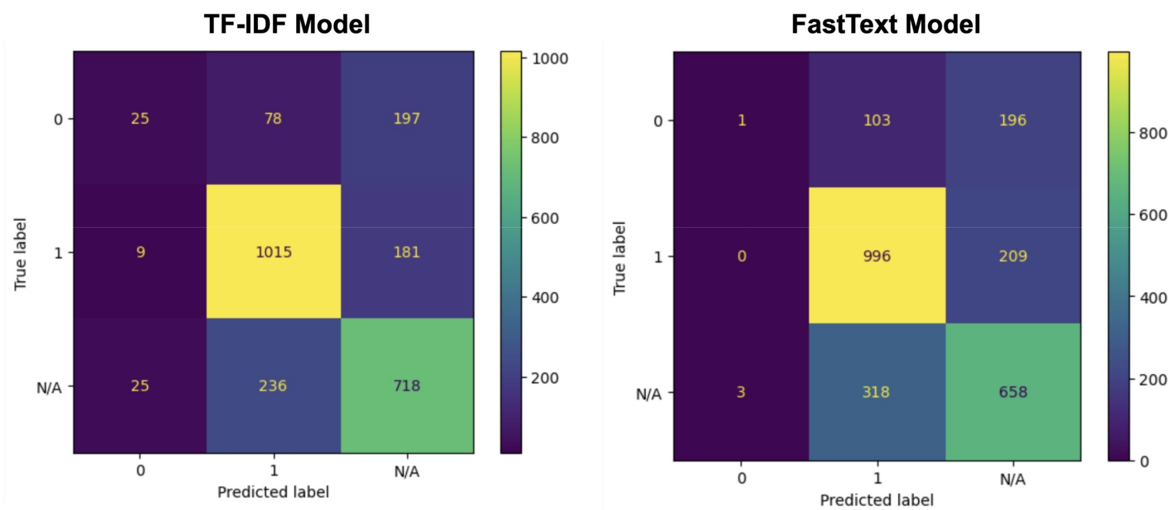

Figure 4.4: Logistic Regression Confusion Matrices

	TF-IDF Model	FastText Model
True Positives	1015 (40.86%)	996 (40.1%)
True Negatives	25 (1.01%)	1 (0.04%)
True N/A	718 (28.9%)	658 (26.49%)
False Positives	314 (12.64%)	421 (16.95%)
False Negatives	34 (1.37%)	3 (0.12%)
False N/A	378 (15.22%)	405 (16.3%)

Table 4.3: Logistic Regression Results

The results in Figure 4.4 indicate that the TF-IDF model correctly predicted 1015 True Positives, 25 True Negatives, and 718 True N/A, while making 314 False Positives, 34 False Negatives, and 378 False N/A predictions. Conversely, the Word2vec or FastText model made 996 True Positives, 1 True Negatives, and 658 True N/A predictions, while incorrectly predicting 421 False Positives, 3 False Negatives, and 405 False N/A.

4.3.3 Naïve Bayes Model

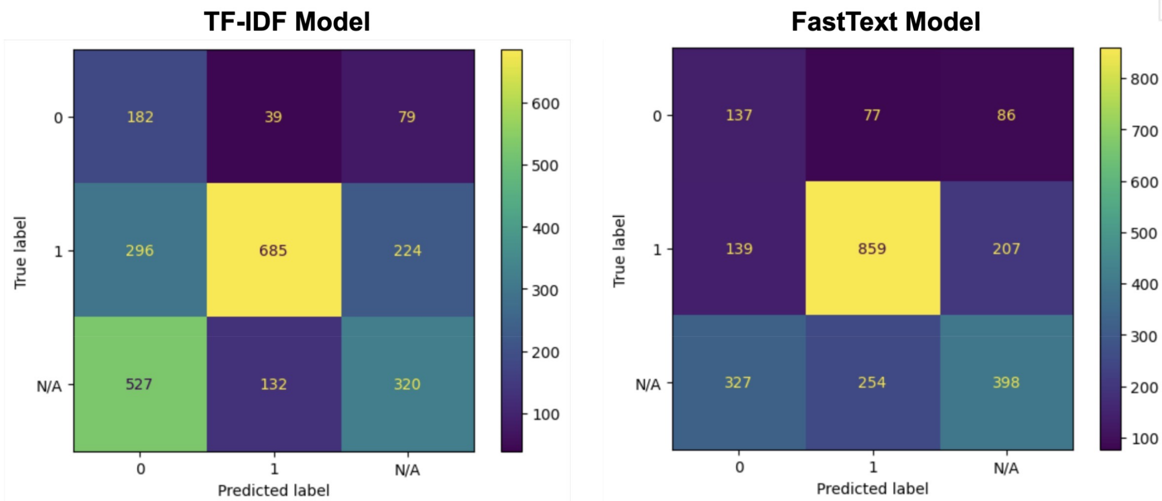

Figure 4.5: Naïve Bayes Confusion Matrices

	TF-IDF Model	FastText Model
True Positives	685 (27.58%)	859 (34.58%)
True Negatives	182 (7.33%)	137 (5.52%)
True N/A	320 (12.88%)	398 (16.02%)
False Positives	171 (6.88%)	331 (13.33%)
False Negatives	823 (33.13%)	466 (18.76%)
False N/A	303 (12.2%)	293 (11.8%)

Table 4.4: Naïve Bayes Results

As illustrated in Figure 4.5, the TF-IDF model showed 658 True Positives, 182 True Negatives, 320 True N/A, 171 False Positives, 823 False Negatives, and 303

False N/A. Conversely, the Word2vec or FastText model yielded 859 True Positives, 137 True Negatives, 398 True N/A, 331 False Positives, 466 False Negatives, and 293 False N/A.

4.3.4 K-Nearest Neighbors Model

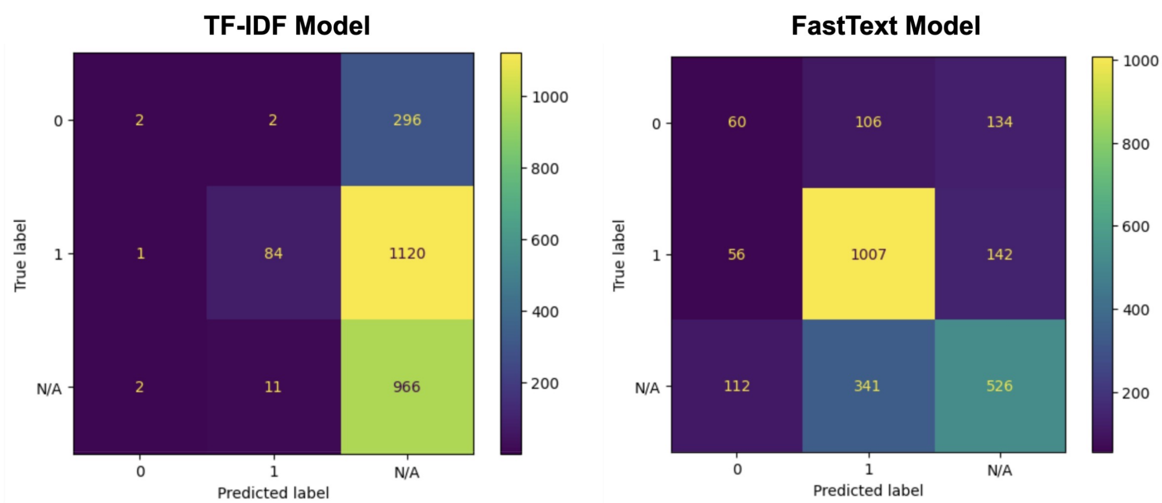

Figure 4.6: K-Nearest Neighbors Confusion Matrices

	TF-IDF Model	FastText Model
True Positives	84 (3.38%)	1007 (40.54%)
True Negatives	2 (0.08%)	60 (2.42%)
True N/A	966 (38.89%)	526 (21.18%)
False Positives	13 (0.52%)	447 (18%)
False Negatives	3 (0.12%)	168 (6.76%)
False N/A	1416 (57%)	276 (11.11%)

Table 4.5: K-Nearest Neighbors Results

The results presented in Figure 4.6 illustrate that the TF-IDF model yielded 84 True Positives, 2 True Negatives, 966 True N/A, 13 False Positives, 3 False Negatives, and 1416 False N/A. Conversely, the Word2vec or FastText model exhibited 1007 True Positives, 60 True Negatives, 526 True N/A, 447 False Positives, 168 False Negatives, and 276 False N/A.

4.3.5 Decision Tree Model

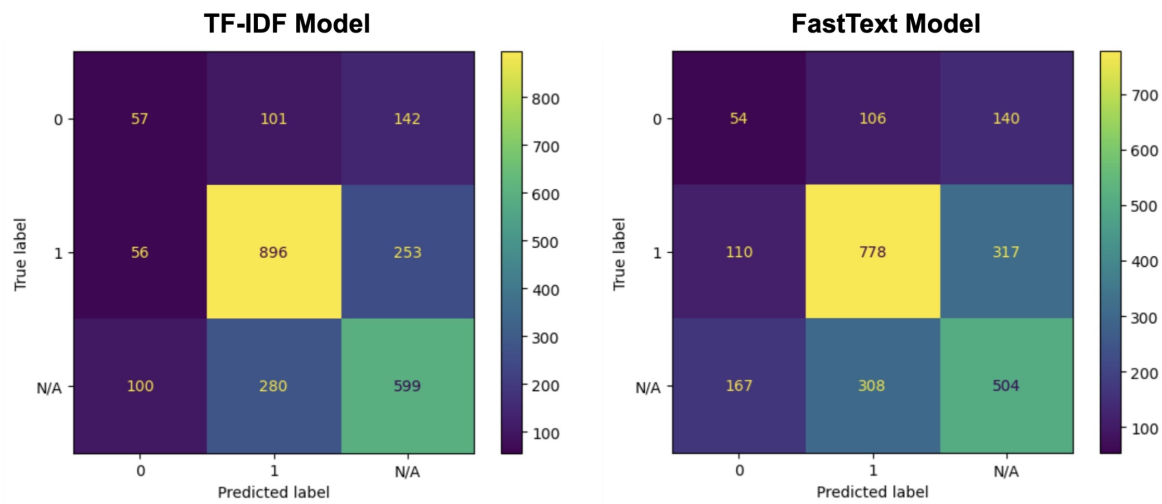

Figure 4.7: Decision Tree Confusion Matrices

	TF-IDF Model	FastText Model
True Positives	896 (36.07%)	778 (31.32%)
True Negatives	57 (2.29%)	54 (2.17%)
True N/A	599 (24.11%)	504 (20.29%)
False Positives	381 (15.34%)	414 (16.67%)
False Negatives	156 (6.28%)	277 (11.15%)
False N/A	395 (15.9%)	457 (18.4%)

Table 4.6: Decision Tree Results

As depicted in Figure 4.7, the TF-IDF model made accurate predictions with 896 True Positives, 57 True Negatives, and 599 True N/A, while committing

381 False Positives, 156 False Negatives, and 395 False N/A. Conversely, the Word2vec or FastText model produced 778 True Positives, 54 True Negatives, and 504 True N/A, with 414 False Positives, 277 False Negatives, and 457 False N/A.

4.3.6 Machine Learning Models

From these findings, aside from the TF-IDF K-Nearest Neighbors Model, the models were able to categorize factual information the best among the three classifications. To check which model performed best, their corresponding metric scores were also calculated.

	Mean Macro F1	Mean Accuracy
TF-IDF Support Vector Machine Model	66% (+/- 3%)	70% (+/- 3%)
FastText Support Vector Machine Model	64% (+/- 2%)	68% (+/- 3%)
TF-IDF Logistic Regression	65% (+/- 3%)	69% (+/- 3%)
FastText Logistic Regression Model	62% (+/- 2%)	66% (+/- 2%)
TF-IDF Naïve Bayes Model	54% (+/- 3%)	53% (+/- 3%)
FastText Naïve Bayes Model	56% (+/- 4%)	55% (+/- 3%)
TF-IDF K-Nearest Neighbors Model	23% (+/- 1%)	40% (+/- 0%)
FastText K-Nearest Neighbors Model	59% (+/- 2%)	61% (+/- 2%)
TF-IDF Decision Tree Model	59% (+/- 2%)	60% (+/- 2%)
FastText Decision Tree Model	51% (+/- 3%)	51% (+/- 3%)

Table 4.7: Machine Learning Models Metric Scores

According to Table 4.7, the TFI-IDF Support Vector Machine Model demonstrated the best performance among all the models, with a mean macro F1 score of 0.66 and a mean accuracy of 0.70, both of which were the highest compared to the other models.

There are several possible reasons why the TF-IDF approach outperformed FastText in this study. One possible reason is that the testing data was included in the vectorization process, enabling the vectorizer to incorporate words that were already present in the dataset as features.

Moreover, previous research has shown that Support Vector Machines (SVM) tend to perform better than other machine learning algorithms in text classification tasks. This is attributed to SVM's capacity to identify essential word features that are predictive of each class and to find a hyperplane that maximally separates the classes. Logistic Regression also performed well, with its straightforward utilization of a line instead of a hyperplane.

4.4 Exploratory Data Analysis

4.4.1 Sentiment of Tweets

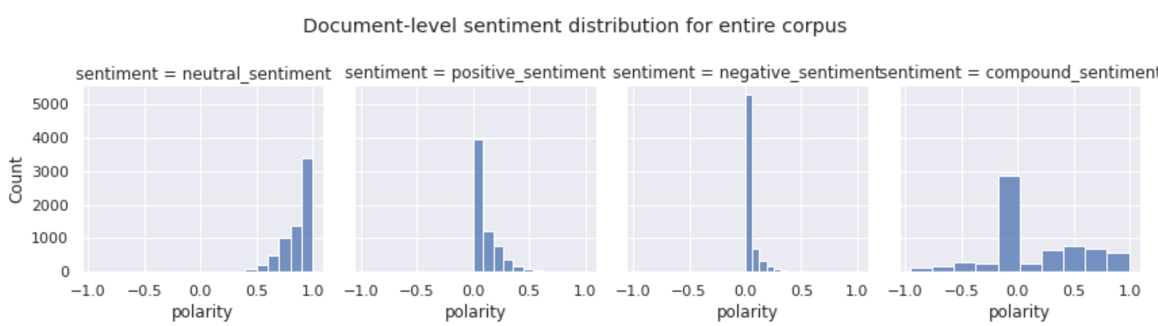

Figure 4.8: Document-level Sentiment of Entire Corpus

The document-level sentiment for the entire corpus leans toward neutral sentiment. The distribution for positive and negative sentiment strongly approaches 0, meaning that most of the corpus is neither overly positive or negative. Interestingly, compound sentiment, although it is mostly neutral, leans slightly towards the right side which could suggest that the tweets show more positive sentiment overall.

4.4.2 Frequency of Terms per Candidate

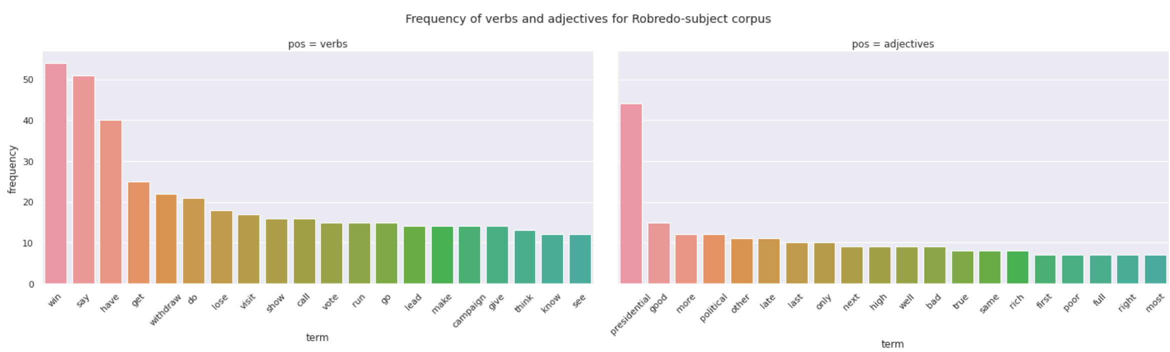

Figure 4.9: Term Frequency for Robredo Tweets

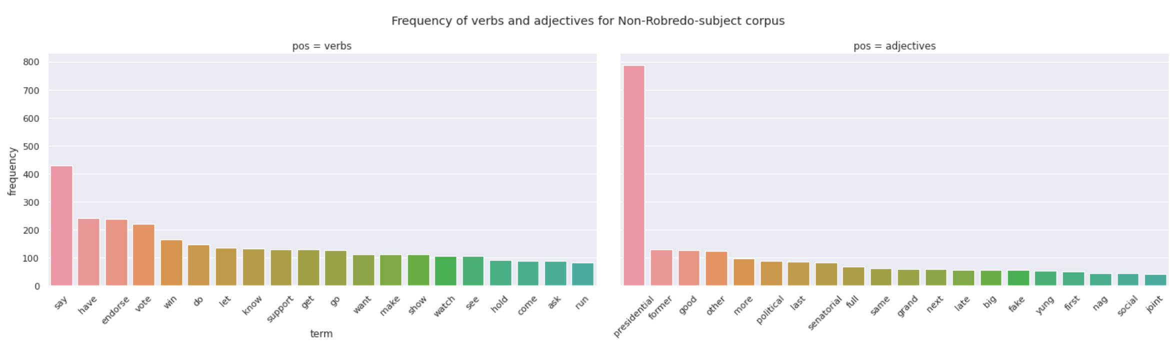

Figure 4.10: Term Frequency for Non-Robredo Tweets

For tweets related to presidential candidate Leni Robredo, it can be seen that the most frequently mentioned verbs center around positive actions about her campaign like “win”, “visit”, “vote”, “run”, and “lead”. However, there are also appearances of negative verbs such as “withdraw” and “lose” which could explain the mixed sentiment of the public towards Robredo. This reflects in the most frequently mentioned adjectives as well which are a combination of both positive and negative terms.

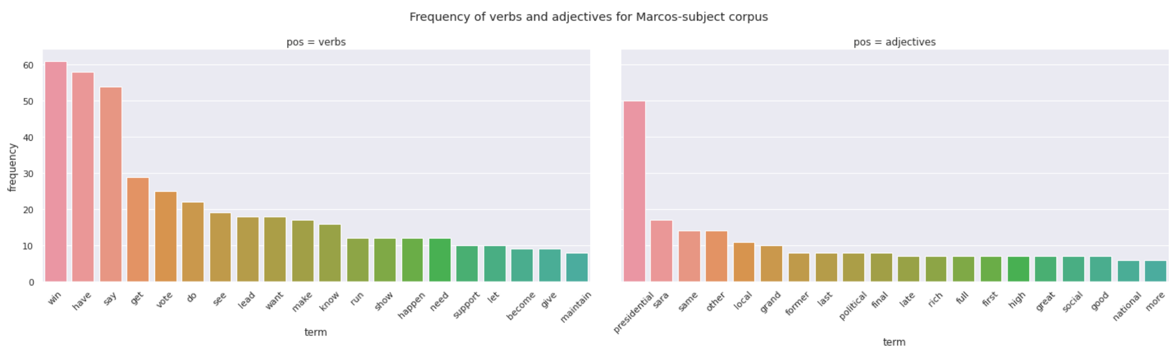

Figure 4.11: Term Frequency for Marcos Tweets

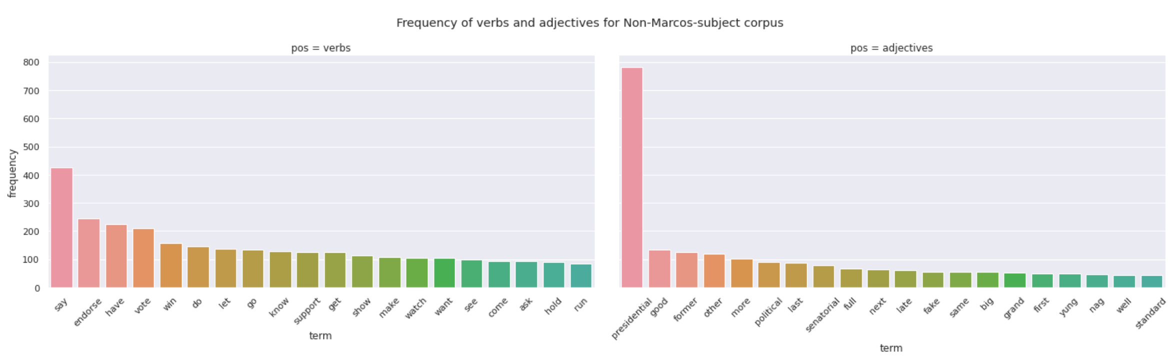

Figure 4.12: Term Frequency for Non-Marcos Tweets

For tweets about presidential candidate Ferdinand Marcos Jr., the most frequently mentioned verbs point towards positive actions about his campaign as well but with the addition of more words that suggest requests such as “need”, “support”, and “maintain”. Notably, while “sara” is not technically an adjective, it is the second most frequently mentioned adjective which, in the context of the presidential election campaign, could point to the vice-presidential aspirant and Marcos’ running mate, Sara Duterte. This could suggest that Marcos and Sara are often talked about as a tandem.

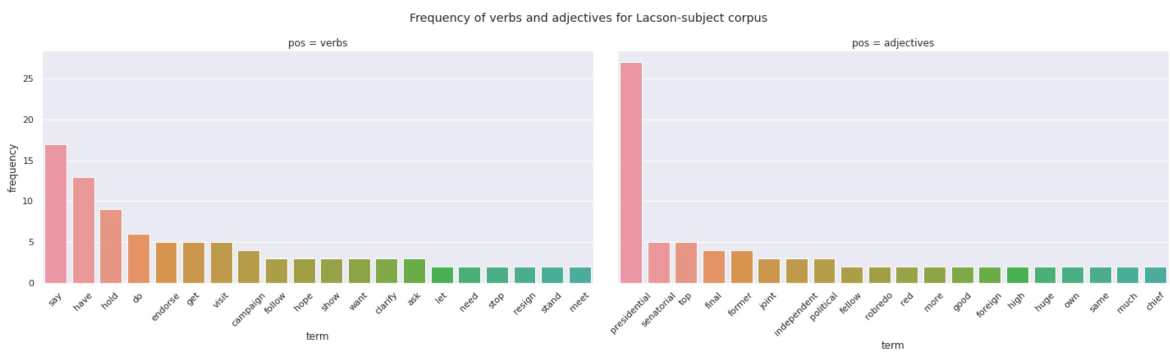

Figure 4.13: Term Frequency for Lacson Tweets

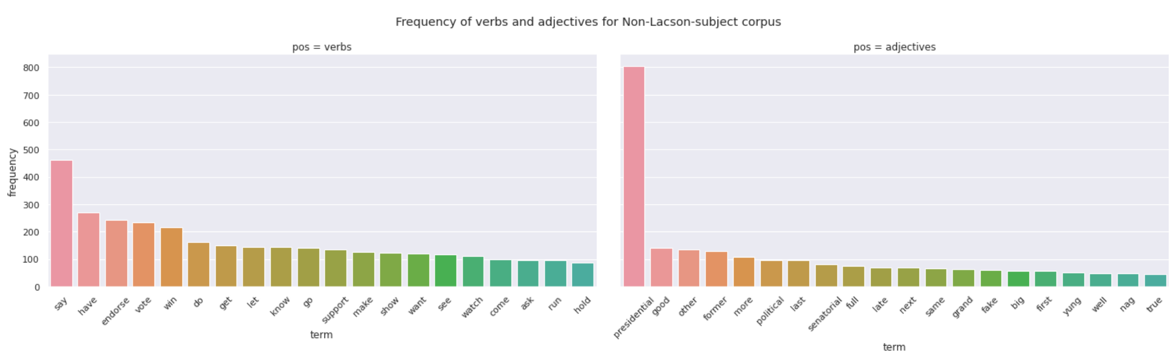

Figure 4.14: Term Frequency for Non-Lacson Tweets

While the tweets about presidential candidate Panfilo Lacson also seem to point towards his campaign, the verb that is most frequently mentioned for Lacson is “say”. This contrasts the most frequently mentioned verb for Robredo and Marcos which is “win”. This could suggest that the public has a less of a positive opinion on Lacson winning the presidential elections. Interestingly, “robredo” also appears in the mentioned adjectives for Lacson which could mean that the two candidates are often compared to each other or talked about together.

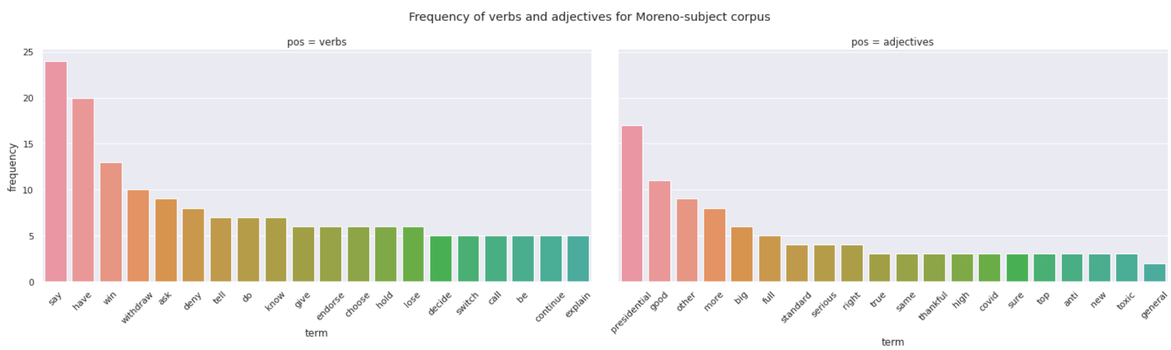

Figure 4.15: Term Frequency for Moreno Tweets

Similar to Lacson, the verb that is most frequently mentioned for Moreno is “say”. However, the difference between the two candidates is shown through the other frequent relevant verbs in the Moreno corpus such as “withdraw” and “switch”, which could pertain to the former candidate’s and his supporters urge for Robredo to withdraw from the presidential race and other supporters to switch to Moreno instead. Negative adjectives, particularly “anti” and “toxic” are also seen in the Moreno corpus, that may indicate his or his supporters' behavior during the campaigning period.

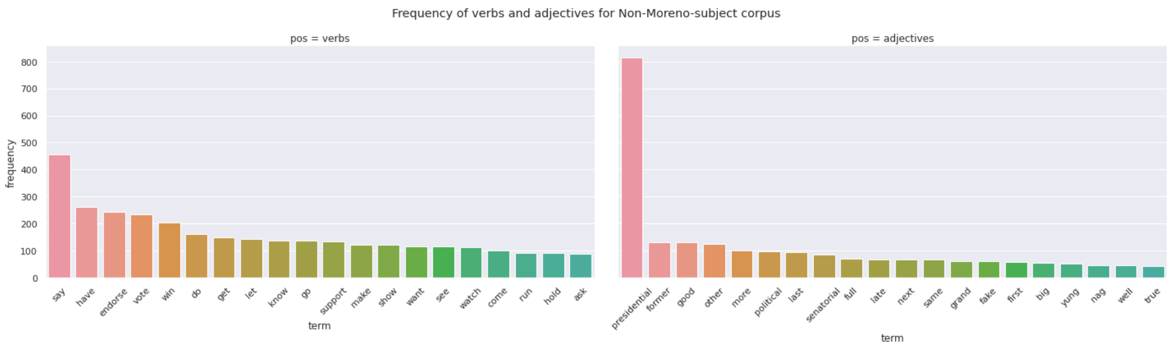

Figure 4.16: Term Frequency for Non-Moreno Tweets

Similar to Lacson and Moreno, the verb that is most frequently mentioned for Pacquiao is also "say". The verbs found in the Pacquaio corpus are common and none seem to have any significant meaning compared to that of the other candidates. What could be inferred from this may be in line with Pacquiao's campaigning with having no strong impact on the media.

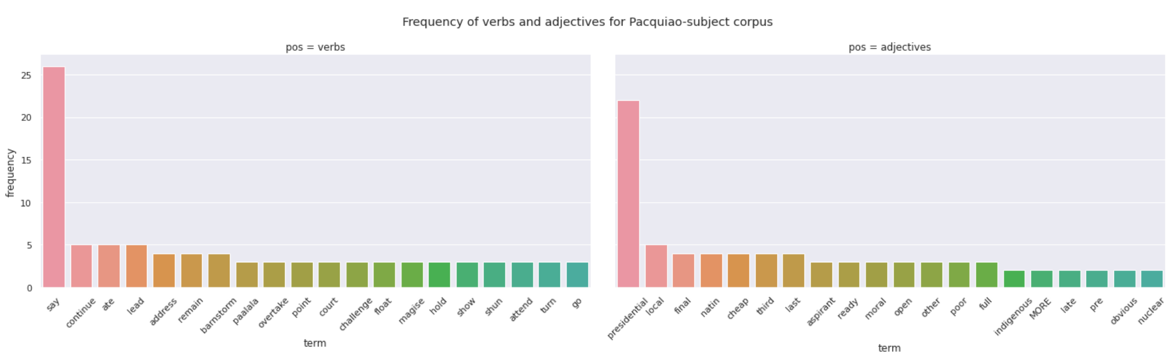

Figure 4.17: Term Frequency for Pacquiao Tweets

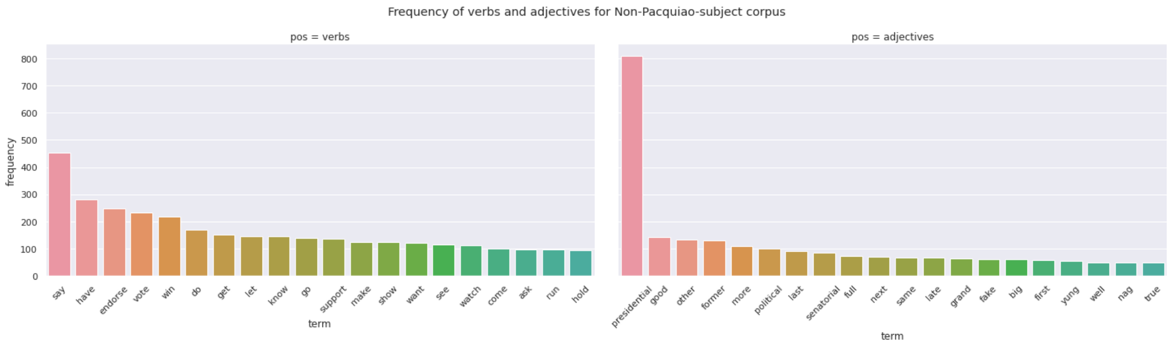

Figure 4.18: Term Frequency for Non-Pacquiao Tweets

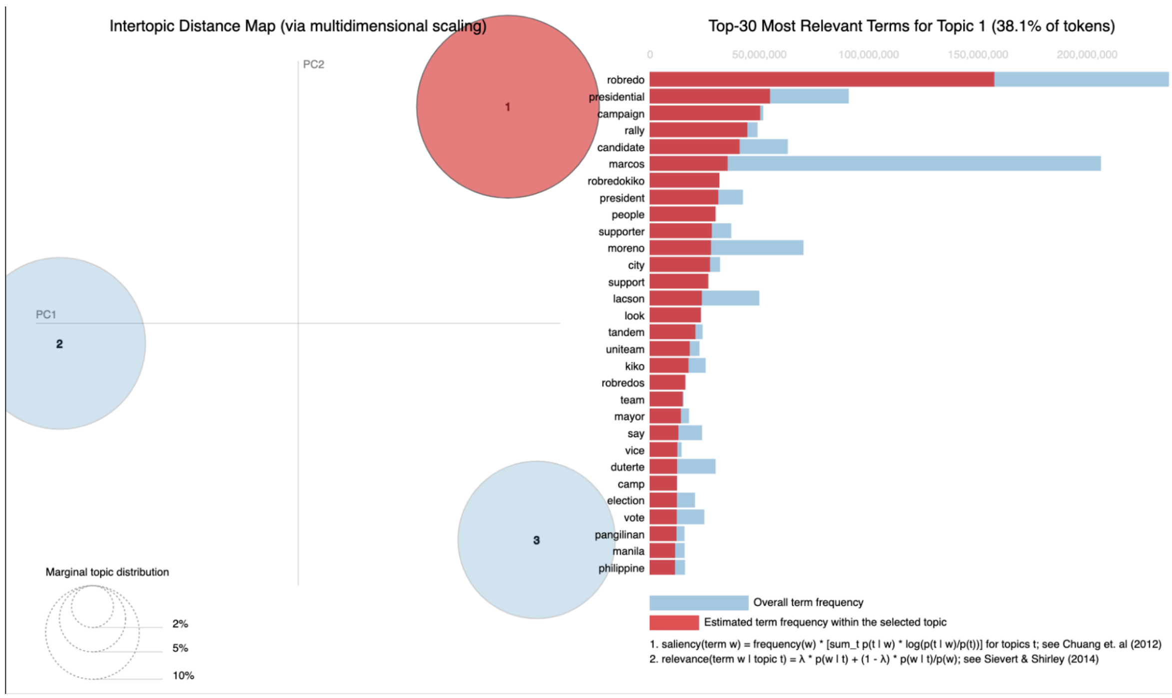

Figure 4.19: Topic Modelling Results 1

The first topic mainly talks about the campaign efforts of the presidential election frontrunners and the support of the people. This is proven by the appearance of the names of the presidential candidates and words that signify collective action such as “campaign”, “rally”, and “supporter”. Interestingly, among all the vice-presidential candidates only the running mate of presidential candidate Leni Robredo, Kiko Pangilinan, appears in the first topic. This could mean that Pangilinan also garners strong opinions and that the tandem is viewed strongly by the public.

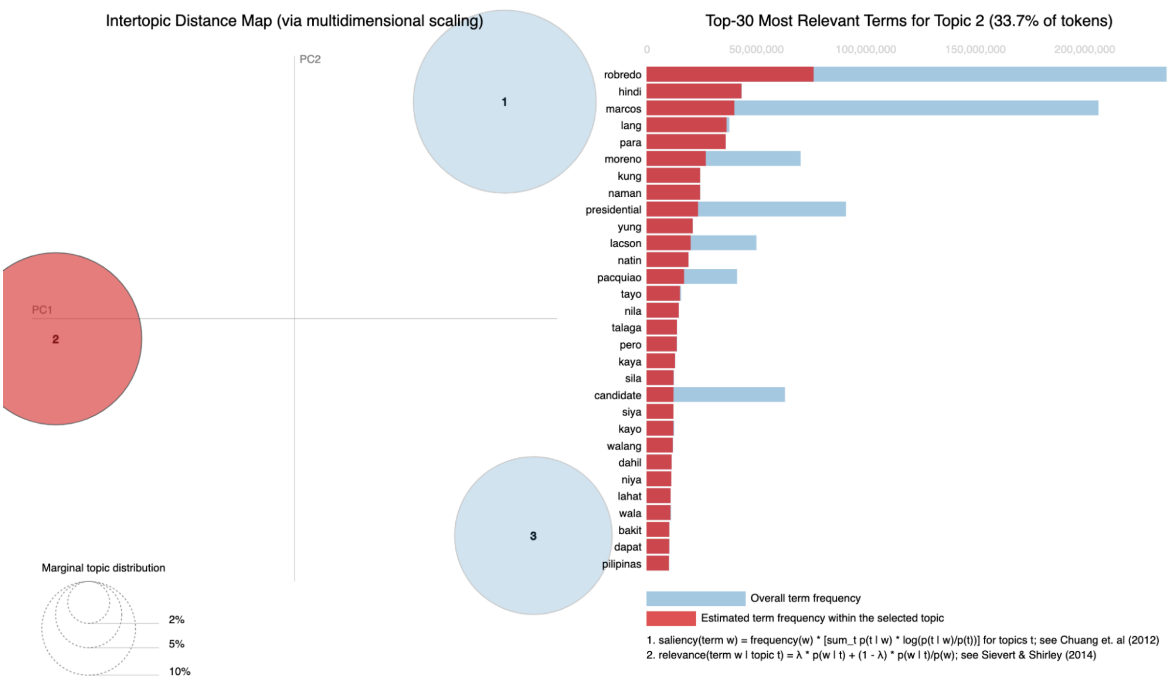

Figure 4.20: Topic Modelling Results 2

Similar to the first topic, the second topic focuses on the general campaign efforts of the presidential candidates but in this topic, it separates the Tagalog terms found in the entire corpus. There does not appear to be any strong sentiment in this topic.

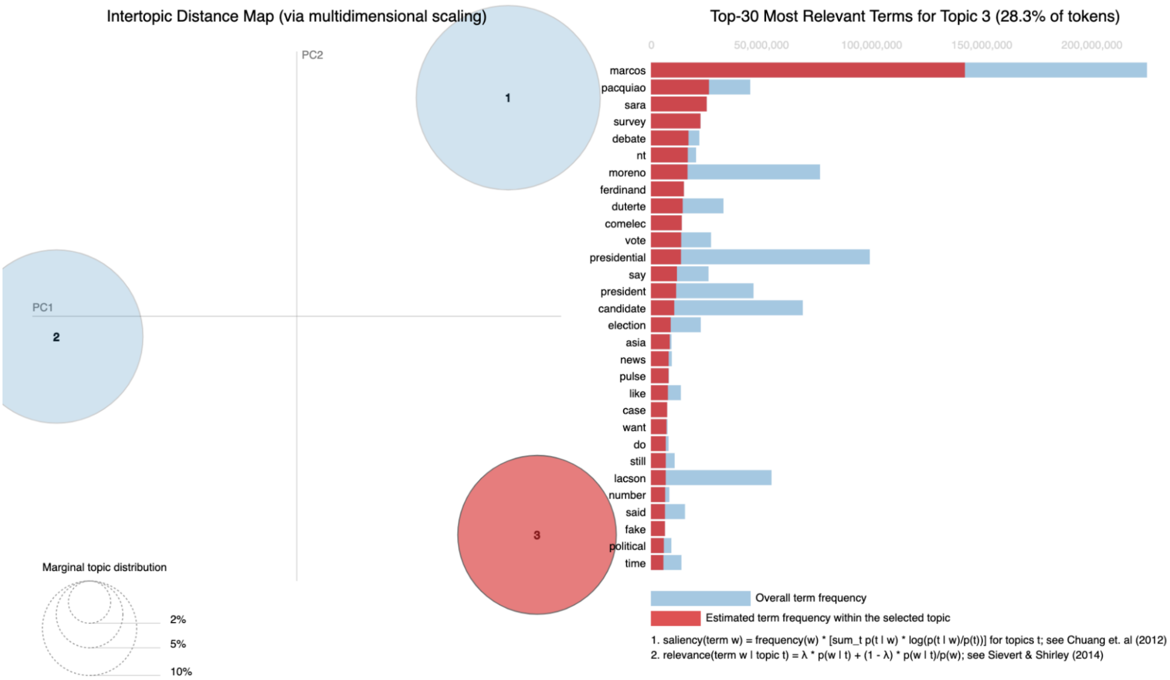

Figure 4.21: Topic Modelling Results 3

The third topic seems to center around the official news reports and surveys during the presidential elections. This is supported by the appearance of words such as "debate", "comelec", "survey", and "news". This cluster may be made up of tweets from the official social media accounts of news organizations and journalists report on these matters without any bias.

CHAPTER V

Conclusions and Recommendations

5.1 Conclusions

5.2 Recommendations

BIBLIOGRAPHY

[1] AGARWAL, N. The ultimate guide to different word embedding techniques in nlp.

[2] AGARWAL, P., HUNT, K., AND ZHUANG, J. *Monitoring Misinformation on Twitter During Crisis Events: A Machine Learning Approach*. ResearchGate, 2020.

[3] APHIWONGSOPHON, S., AND CHONGSTITVATANA, P. *Identifying misinformation on Twitter with a support vector machine*. ResearchGate, 2020.

[4] BRADSHAW, S., AND HOWARD, P. N. *The Global Disinformation Order 2019 Global Inventory of Organised Social Media Manipulation*. Computational Propaganda Research Project, 2019.

[5] CABRAL, L., MONTEIRO, J. M., DA SILVA, J. W. F., MATTOS, C. L., AND , P. J. C. M. Fakewhastapp.br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. *CoRR 1* (2021).

[6] CHOUDRIE, J., BANERJEE, S., KOTECHA, K., WALAMBE, R., KARENDE, H., AND AMETA, J. Machine learning techniques and older adults process-

ing of online information and misinformation: A covid 19 study. *Computers in Human Behavior 119* (2021), 106716.

[7] IBM. What is exploratory data analysis?

[8] JANG, S. M., GENG, T., LI, J.-Y. Q., XIA, R., HUANG, C.-T., KIM, H., AND TANG, J. A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior 84* (2018), 103–113.

[9] JIN, Z., CAO, J., GUO, H., ZHANG, Y., WANG, Y., AND LUO, J. Detection and analysis of 2016 us presidential election related rumors on twitter. In *Social, Cultural, and Behavioral Modeling* (Cham, 2017), D. Lee, Y.-R. Lin, N. Osgood, and R. Thomson, Eds., Springer International Publishing, pp. 14–24.

[10] KAUR, K., NAIR, S., KWOK, Y., KAJIMOTO, M., CHUA, Y. T., LABISTE, M. D., SOON, C., JO, H., LIN, L., LE, T. T., AND KRUGER, A. *Information Disorder in Asia and the Pacific: Overview of Misinformation Ecosystem in Australia, India, Indonesia, Japan, the Philippines, Singapore, South Korea, Taiwan, and Vietnam*. University of Hong Kong, 2018.

[11] MATHWORKS. What is word2vec?

[12] NEWMAN, N., FLETCHER, R., KALOGEROPOULUS, A., AND NIELSEN, R. Digital news report 2019, 2019.

[13] NEYAZI, T., KALOGEROPOULOS, A., AND NIELSEN, R. Misinformation concerns and online news participation among internet users in india.

[14] NICHOLSON, C. A beginner's guide to word2vec and neural word embeddings.

[15] NIKHILAGGARWAL3. What is exploratory data analysis?

[16] OGNYANOVA, K., LAZER, D., ROBERTSON, R., AND WILSON, C. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School (HKS) Misinformation Review* (2019).

[17] OGNYANOVA, K., LAZER, D., ROBERTSON, R., AND WILSON, C. Global social media statistics, 2022.

[18] PATIL, P. What is exploratory data analysis?

[19] RITURAJSAHA. Understanding tf-idf (term frequency-inverse document frequency).

[20] SCIKIT LEARN. sklearn.feature$_{e}xtraction.text.tfidfvectorizer.$

[21] SIMHA, A. Understanding tf-idf for machine learning.

[22] STARMER, J. *StatQuest: K-nearest neighbors, Clearly Explained*, 2017.

[23] STARMER, J. *StatQuest: Logistic Regression*, 2018.

[24] STARMER, J. *StatQuest: Random Forests Part 1 - Building, Using and Evaluating*, 2018.

[25] STARMER, J. *Support Vector Machines Part 1 (of 3): Main Ideas!!!*, 2019.

[26] STARMER, J. *Naive Bayes, Clearly Explained!!!*, 2020.

[27] STECANELLA, B. Understanding tf-idf: A simple introduction.

[28] SU, Q., WAN, M., LIU, X., AND HUANG, C.-R. Motivations, methods and metrics of misinformation detection: An nlp perspective. *Natural Language Processing Research 1* (2020), 1–13.

[29] TENSORFLOW. word2vec.

[30] UYHENG, J., AND CARLEY, K. M. Bot impacts on public sentiment and community structures: Comparative analysis of three elections in the asia-pacific. In *SBP-BRiMS 2020: Social, Cultural, and Behavioral Modeling* (2020), Springer.

[31] VOSOUGHI, S., ROY, D., AND ARAL, S. The spread of true and false news online.