FROM POSTS TO VOTES: USING RoBERTa TO ANALYZE SOCIAL MEDIA AS AN INDICATOR OF PRESIDENTIAL ELECTION WINS FOR THE PHILIPPINES AND UNITED STATES

A Thesis

Presented to the

Department of Information Systems

and Computer Science

Ateneo de Manila University

In Partial Fulfillment

of the Requirements for the Degree of

Bachelor of Science in Computer Science

by

Michael Gavin N. Del Castillo,

Mary June Aubrey C. San Jose,

Shaira Jane L. Sy

2026

ABSTRACT

Recent election periods have seen the rise of social media platforms as tools through which political ideologies are spread, and support for candidates are garnered. Despite widespread support in online spaces, however, former vice presidents Maria Leonor "Leni" Robredo and Kamala Devi Harris lost their respective presidential elections in 2022 (in the Philippines) and 2024 (in the United States) to Ferdinand “Bongbong” Marcos Jr. and Donald John Trump. The main motivation for this research is to determine the effectiveness of social media as an indicator of electoral wins. The research will make a similar comparison between the 2016 Philippine presidential elections with leading candidates Rodrigo Duterte and Manuel “Mar” Araneta Roxas II, and the 2020 US presidential elections, with leading candidates Donald John Trump and Joseph Robinette “Joe” Biden Jr. Through the Robustly Optimized BERT Pre-Trained Approach (RoBERTa) and other NLP techniques, this paper aims to perform sentiment analysis on posts from Facebook, TikTok, and X (formerly Twitter) to draw insights on how citizens perceived both the winning and first runner-up electoral candidates during pre- and proper election periods and compare these perceptions to the results of each presidential election.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Recent election periods have seen the rise of social media platforms as tools through which political ideologies are spread, and support for candidates are garnered. Despite widespread support in online spaces, however, former vice presidents Maria Leonor "Leni" Robredo and Kamala Devi Harris lost their respective presidential elections in 2022 (in the Philippines) and 2024 (in the United States) to Ferdinand “Bongbong” Marcos Jr. and Donald John Trump. The main motivation for this research is to determine the effectiveness of social media as an indicator of electoral wins. The research will make a similar comparison between the 2016 Philippine presidential elections with leading candidates Rodrigo Duterte and Manuel “Mar” Araneta Roxas II, and the 2020 US presidential elections, with leading candidates Donald John Trump and Joseph Robinette “Joe” Biden Jr. Through the Robustly Optimized BERT Pre-Trained Approach (RoBERTa) and other NLP techniques, this paper aims to perform sentiment analysis on posts from Facebook, TikTok, and X (formerly Twitter) to draw insights on how citizens perceived both the winning and first runner-up electoral candidates during pre- and proper election periods and compare these perceptions to the results of each presidential election.

CHAPTER I

INTRODUCTION

1.1 Context of Study

In recent decades, social media has had a major role as a platform through which political ideologies are spread and political discussions occur. This is especially apparent when observing the flow of recent elections in certain countries. In the Philippines, the 2016 Philippine presidential election is widely considered the first "social media election" in the Philippines, mainly due to how its winner, Rodrigo Roa Duterte, was able to utilize social media to establish a controversial image which mobilized his wide follower-base to rally in support of him, both online and offline [36]. Meanwhile, social media proved to be a crucial element in Joseph Robinette "Joe" Biden Jr.'s 2020 election win in the United States. Through an "influencer campaign", Biden was able to reach out to young audiences on social media, particularly those of generation Z, which then translated to a massive voter turnout in that certain demographic [37].

In other elections, however, cases have occurred in which online support did not directly translate to election wins. One of the most prominent examples of the importance of social media in world politics is former United States (US) Vice

President Kamala Harris' social media campaign for the 2024 US presidential elections. Mainly targeting younger audiences through viral "memes" and other social media trends, Harris was able to amass widespread support for her campaign, having just over 5 million followers supporting her endeavors on TikTok and X (formerly Twitter) combined [25]. Similarly, the 2022 Philippine elections saw the Angat Buhay campaign of former Vice President Leni Robredo. Similarly to Harris, Robredo was able to garner the attention of young audiences on social media. Rallies in support of Robredo alongside Robredo's track record as a politician made her a popular choice for millions as a capable presidential candidate [22].

Despite massive online support, both Harris and Robredo had lost their respective elections, the former only garnering 226 electoral votes (against Donald Trump's 312 votes) and the latter garnering some 14.8 million votes (as opposed to Ferdinand "Bongbong" Marcos, Jr, who gathered 31.1 million) [2, 8]. Given a possible disparity between social media popularity and election votes, the aim of this research is to provide a data-driven analysis on the effectiveness of social media as an indicator of election wins by observing social media trends at the time of both 2022 and 2024 elections, as well as comparing and contrasting these elections in terms of said trends.

Through Natural Language Processing (NLP) Algorithms, this paper

aims to analyze the conversations on the aforementioned presidential candidates that had transpired in online spaces during pre-election seasons, namely Facebook, TikTok, and X (formerly Twitter). Previous research endeavors have already shown the effectiveness of sentiment analysis in determining key themes behind social media posts, especially in the context of events such as elections. Thus, this research would like to push this idea further by not only contextualizing the data within a single setting. Rather, this paper aims to compare and contrast the election periods of the Philippines and US, given that, as already mentioned, the two countries experienced a supposed upset in terms of electoral candidate votes relative to their presence on social media.

To provide a more thorough analysis, this paper also intends to perform the same analysis on social media spaces during the 2016 elections for both the Philippines and US. Achieved through a comparative study between: (1) the 2022 Philippine presidential election campaigns of candidates Ferdinand “Bongbong” Marcos and Leni Robredo, with their respective running mates Sara Duterte and Francis “Kiko” Pangilinan, and the 2024 US presidential election campaigns of candidates Donald Trump and Kamala Harris, with their respective running mates James “JD” Vance and Timothy Walz; and (2) the 2016 Philippine presidential election campaigns of candidates Rodrigo Duterte and Manuel “Mar” Roxas, with their respective running mates Alan Cayetano and Leni Robredo,

and the 2020 US presidential election campaigns of candidates Joe Biden and Donald Trump, with their respective running mates Kamala Harris and Mike Pence. This research aims to determine whether or not social media support directly translates to election success, or if other factors were present which had contributed to the losses of Harris and Robredo in their respective runs for presidency.

1.2 Research Questions

1. Natural Language Processing (NLP) techniques can be used for sentiment analysis on social media posts. Can these techniques, specifically the RoBERTa-model, be used to predict election outcomes?

    (a) Is RoBERTa able to accurately capture the texts and sentiments shared by Philippine social media users from US social media users?

    (b) Do the different themes, frequencies, and sentiments of keywords and phrases expressed by users on Facebook, X, (formerly Twitter), and TikTok indicate their support for the candidates?

    (c) Can a developed visualization effectively illustrate social media presences throughout the four elections (Marcos Jr. vs. Robredo, Duterte vs. Roxas in the 2022 and 2016 Philippine elections respectively; and

Trump vs. Harris, Trump vs. Biden in 2024 and 2020 US elections respectively), and whether or not these presences predicted their electoral results?

(d) Do the results from the RoBERTa model for the Philippines and United States have the same predictions and findings or do they contradict?

1.3 Research Objectives

1. Natural Language Processing (NLP) techniques can be used for sentiment analysis on social media posts. For this paper, NLP and RoBERTa are being tested on its ability to predict election outcomes.

   (a) Determine the different texts on social media platforms to draw insights on how different presidential candidates were perceived by social media users.

   (b) Analyze different texts on social media platforms to draw insights on how different presidential candidates were perceived by social media users, how often these perceptions are shared, and the general sentiment.

   (c) Develop a dashboard providing a comprehensive overview of social media sentiments in relation to election results by comparing results

drawn from sentiment analysis algorithms and actual elections results

(d) Compare the conclusions of the first two objectives with one another to reaffirm the consistency and accuracy of the algorithm's output for all 4 analyzed elections.

1.4 Scope and Limitations of the Study

This paper consists of two case studies to provide a rich analysis of social media data extracted from four presidential election periods. To add to the discussion, social media data about each president's respective vice presidential running mate candidate will be considered as well. As such, first, a case study analysis comparing the recent 2022 Philippine presidential election, with leading candidates Marcos-Duterte and Robredo-Pangilinan against the recent 2024 US presidential election, with leading candidates Trump-Vance and Harris-Walz. Second, a comparison between the 2016 Philippine presidential election, with leading candidates Duterte-Cayetano and Roxas-Robredo and the 2020 US presidential elections, with leading candidates Trump-Pence and Biden-Harris. The paper aims to analyze how the respective campaign periods of each candidate (and their running mates) were reflected on different social media spaces. This is to determine the effectiveness of social media platforms as indicators of electoral wins and determine whether electoral results can be foreseen based on

online traction and popularity.

This paper will focus only on the social media platforms Facebook, X (formerly Twitter), TikTok, in data collection due to their popularity within the United States and the Philippines. Only posts made after the announcement of a candidate's intention to run for president and prior to the actual election days will be collected, as the goal is to compare pre-election social media data to post-election results. As such, with reference to the dates of when each leading candidates announced their intention to run, the paper will consider posts made after November 21, 2015 for Duterte and July 31, 2015 for Roxas in the 2016 Philippine presidential election, October 7, 2021 for Robredo and October 5, 2021 for Marcos in the 2022 Philippine presidential election, April 25, 2019 for Biden and January 21, 2017 for Trump in the 2020 US presidential election, and July 21, 2024 for Harris and November 15, 2022 for Trump in the 2024 US presidential election [34, 13, 24, 10, 9, 17, 39, 30]. Posts made 2 days before the election dates and beyond will be excluded from the study. It is also worth mentioning that, for all elections under the scope of this study, only the winner and first-runner up will be considered as collecting enough data on all candidates might not be feasible [28]. Finally, *"posts"* include any and all publicly available posts made by the general public on the selected candidates alongside any posts made by the candidates themselves; however, data on the users who made the posts themselves,

such as gender or location, will not be considered due to the lack of availability.

Finally, attached to the names of some candidates are certain criminal cases. The Marcos Family was responsible for a series of atrocities and human rights violations in the 1970s, and Trump is currently facing multiple criminal cases [21, 20]. The paper will not explore such topics in-depth as they are outside of the scope and focus of the study; however, these may be touched upon briefly if it is a popular discussion point among users in the data collected.

1.5 Significance of the Study

This paper aims to contribute to the field of social computing by analyzing the behavior of users in online spaces during election periods. By way of sentiment analysis and other NLP techniques, the research aims to gather statistically large amounts of social media data that represent the general public, and after which makes use of models that process such data and draw insights from the sentiments behind social media posts concerned with presidential elections.

The analytics and findings of this study will benefit the general public's knowledge of how social media runs in both Philippine and American contexts, especially during the election seasons. These findings will also provide them with more context between the spheres of social media in terms of various political ideologies the said candidates have. The researchers hope to impart realizations

on how social media might not be the definitive factor behind electoral wins.

CHAPTER II

REVIEW OF RELATED LITERATURE

The review of related literature of the research examines existing studies grouped according to the following discussion points: the importance of sentimental analysis, especially in the context of presidential elections; the definition of the BERT model and how it is effective in classifying sentiments of social media posts; how social media has become a critical tool for political engagement and in building the general public's sentiment; and backgrounds on the Philippine and US presidential elections that are to be analyzed.

The sentiment analysis subsection discusses its definition and how it is essential in analyzing social media engagement. Next, tools used for analyses will be discussed, especially the transformation model `Bidirectional Encoder Representations from Transformers (BERT)`— its architecture, a variation of RoBERTa, and how effective BERT is for sentiment analysis using various studies as evidence. Lastly, the Elections and Social Media subsection discusses how social media shapes the general public, especially in the context of the Philippines and the United States.

2.1 Sentiment Analysis for Social Media and Elections

According to Liu [2012], the study of people's views, sentiments, assessments, appraisals, attitudes, and emotions about goods, services, organizations, people, problems, events, subjects, and their characteristics is commonly referred to as sentiment analysis or opinion mining [26]. With the explosive growth of social media, it has become a hotspot of opinions, shaping our decisions, especially in an important political event like elections. In the field of social computing, election seasons are one of the widely researched topics, especially on how the interaction in social media affects society in terms of making decisions on who to vote for.

There are examples of studies using sentiment analysis to analyze social media activity. In a study by Macrohon, et al. [2022] and Demillo, et al. [2023], they used the `Naïve Bayes classifier`, a probabilistic learning method, to determine the probability of a tweet belonging to the best class—applicable in determining the polarity of a post [28, 14]. Then, previous studies showed the usage of bidirectional encoder representation from transformers (BERT) models, modified to handle emojis and Tagalog language tweets. Aquino, et al [2025] introduced the emotion-infused `BERT-GCN model` for sentiment analysis, which includes emoji semantics into the models, treating them as sentiment represen-

tation [6]; meanwhile, Cruz, et al. [2022] used the `RoBERTa-tagalog-cased model` to get the vectorized version of Tagalog embeddings, essential to map echo chambers on Twitter via `K-Means modeling` [12]. Lastly, the `Support Vector Machines (SVM) Classifier model` was used by Demillo, et al. [2023] to handle binary classification of data, classifying them as either a negative or positive sentiment [14].

For reasons discussed more in-depth in the next section, BERT was chosen for the study's methodology given its capabilities of performing nuanced analyses and classification of social media posts.

2.1.1 BERT and RoBERTa

Recent developments in devising models for NLP tasks have ensured that models are updated to be more context-aware, being able to provide a more holistic and nuanced analysis of certain texts. One such model is BERT, short for Bidirectional Encoder Representations from Transformers (referred to henceforth as BERT). Developed by the Google AI Language Laboratory, the main advantage provided by BERT is its ability to analyse text in a bidirectional manner, as opposed to more traditional machine learning models, such as GPT, which only analyse text left-to-right or vice versa. Bidirectional analyses of text ensures that BERT is able to capture not only the sentiments of text, but do so in such a manner that the model is able to detect certain nuances, such as sarcasm or

irony. BERT's architecture is built on transformers, an architecture of neural networks that uses a combination of recurrent and convolutional networks.

BERT's architecture is built on transformers, an architecture of neural networks that uses a combination of recurrent and convolutional networks.

BERT is primarily pre-trained in two phases: first, using a large dataset of unlabeled data, then second; a smaller set of labeled data, usually for fine-tuning the BERT model according to some NLP task. One such NLP task is Sentiment Analysis [23]. In pretraining, BERT operates on two main objectives. The first is the Masked Language Model (MRM henceforth). A random sample of tokens in the input sequence is selected and replaced with a special mask token [MASK]. The objective is then for the BERT model to be able to predict what these masked tokens are. Next is Next Sentence Prediction (NSP), a binary classification task. The goal is for the model to be able to predict whether two text segments follow each other. Both positive examples (consecutive sentences from the training set) and negative examples (pairs of segments from different documents) are provided and are sampled with equal probability.

In 2019, the Facebook AI research team found that BERT was "significantly undertrained", and thus proposed RoBERTa, short for "[A] Robustly Optimized BERT Pretraining Approach". The team sought to improve the training process of the BERT model by (1) training the model over a longer period of time,

and with bigger batches of data, (2) removing the NSP objective in pretraining, and (3) dynamically applying the masking pattern applied to the training data. The results of this optimized pretraining process do show, indeed, that RoBERTa is able to either match or exceed the performance of BERT in NLP tasks, the former scoring higher than the latter in multiple NLP model evaluation tests such as GLUE, SQuAD, and RACE [27].

As BERT and RoBERTa have seen usage in analysing large datasets of text, it is able to aid in research on social media. Social media is considered a rapidly evolving form of text widely different from more traditional text formats such as novels mainly due to the widespread usage of informal language, abbreviations, and emojis, among other elements, which can be challenging to understand without the proper context.

For one, Kumar and Sadanandam [2023] were able to use BERT and RoBERTa to classify a large dataset of some 8,225 tweets related to the Coronavirus into three general sentiments: positive, neutral, and negative. Both BERT and RoBERTa were able to perform sentiment analysis across the entire dataset, achieving high accuracies (at least 88%), precision (at least 0.88), recall (at least 0.74 but can go as high as 0.91), and F1-score (at least 0.78 but can go as high as 0.90) [32]. Prasanthi, et al. [2023] were also able to accomplish a similar feat using both BERT and RoBERTa, performing sentiment analysis on large social

media datasets with extremely high accuracies, these accuracies only improving with each succeeding epoch. BERT was able to achieve a base accuracy of 95.10% on the first epoch, which only increased to 99.16% on the tenth epoch. Similarly, RoBERTa was able to achieve a base accuracy of 99.53%, with a final accuracy of 99.70% on the tenth epoch [33].

As mentioned earlier, BERT and RoBERTa are able to capture nuances such as sarcasm and irony in texts. Detecting sarcasm, in particular, has proven to be a highly challenging NLP task as a sarcastic statement implies a negative sentiment whilst seemingly conveying a positive one surface-level. Nevertheless, Dong, et al. [2020] were able to train RoBERTa on a dataset of posts from Reddit and X (formerly Twitter) to give it the ability to detect sarcasm in a given text, with an F1 score of 80.2 [15].

These studies illustrate the importance of RoBERTa and BERT in the context of sentiment analysis on highly informal and nuance-laden texts such as social media posts.

2.2 Social Media Use and the Elections

2.2.1 Candidate Activity

The 2016 Philippine presidential election is widely considered the first "social media election" in the Philippines [36]. The two Philippine presidential elections, 2016 and 2022, are undoubtedly linked as they involve something other than the rise of the Marcos-Duterte alliance: the prominence of social media as a means to bolster their presence in people's lives and boost their popularity.

Despite having the most engagement, Duterte's online presence during his presidential campaign is nothing short of lackluster and underwhelming [36].

On the other hand, Marcos Jr's campaign has been well-established and maintained in the years leading up to his campaign [29]. His pitch throughout the campaign calls for national unity, featuring the glorification of his father's legacy. In Rappler's three-part study on "networked propaganda" back in 2019, there was a rise in many pro-Marcos pages and channels on different social media platforms, notably on TikTok. There was less activity from Marcos Jr. himself, however, those channels were particularly full of pro-Marcos content [29].

2.2.2 Public Opinion

Duterte's successful campaign can be attributed to his aggressive supporters– most of whom are vocal online and active offline. As observed by Sinpeng, et al. [2020], despite Duterte's unprofessional online presence, his supporters are committed and constantly rallied to his defense against the criticism of other candidates [36]. There are also prospects of the heavy involvement of informal actors like paid trolls and influencers as having major roles in mobilizing (and agitating) digital communities, which helped spread his popularity [36].

"The recent election has been the most social media-active and engaging campaign in the country's democratic history." said Ampon, et al. [2023] in a paper analyzing the political message strategies of Marcos and Robredo. In the 2022 Philippine presidential race, both leading candidates (Marcos Jr. and Robredo) have taken great leads on social platforms like Facebook and X, respectively [4]. Electoral campaigns are aimed at spreading awareness about the candidate's identity and, over the years, Marcos Jr. has amassed a large number of supporters on TikTok based on the top 4 trending hashtags related to him: `#bongbongmarcos` (3.4 billion views), `#bbmsara2022` (2.3 billion views), `#uniteam` (2.5 billion views) and `#bbm2022` (2 billion views) [29].

2.3 Elections Background

2.3.1 Philippine Elections (2016, 2022)

In his study on the 2016 Philippine presidential elections, after analyzing pre-election surveys, the candidates' campaign strategies and their advocacies, and their supporters' age demographic and news tracking, Holmes [2016] observed that the elections in the Philippines are political clan-dominated, personality-oriented, and media driven [19].

Rodrigo Duterte's victory in the election was believable to the public and was attributed to: the clarity of his campaign slogan, his significant support from a geographic area, and how he criticized and questioned the character and competence of his fellow candidates. However, one of the most significant observations in the study is the importance of the media, which is updated in real-time and where voter preference was shaped and reformed by Duterte's critiques and 'bashing' [19].

Following Duterte's term in office, Ferdinand 'Bongbong' Marcos Jr.'s electoral win has caused much uproar among the nation's scholars. The general resurgence of the Marcos Clan in politics can be attributed to 3 factors: (1) the people's nostalgia of the Marcos era, (2) Duterte's political influence, and (3) the Marcos' years-long digital disinformation campaign on social media [31].

Duterte's consequent influence on Marcos' resurgence cannot be dismissed, as signs have pointed out that Duterte's indirect endorsement led to his win [16].

2.3.2 US Elections (2020, 2024)

At the cusp of the COVID-19 pandemic, the 2020 US presidential election had taken a major hit– particularly for one of its leading candidates, Donald Trump, whose vote share is largely affected by COVID-19-related cases [7]. It is likely that Trump was viewed negatively for how he handled the pandemic as the most affected counties and states are ones without stay-at-home orders, in swing states, or states that Trump won in 2016. This mismanagement is what likely led to changes in voter preferences and Joe Biden's eventual electoral win.

The 2024 US presidential election was predicted to be one of the most competitive in modern history with a tight competition between candidates Donald Trump and Kamala Harris, the new face of the Democratic Party [35]. In the end, Trump had managed to win the electoral race [35].

2.4 Data Collection Methods

In the Philippine context, as Filipino is a low-resource language, a lack of Filipino datasets of tweets has been a pressing issue from previous studies [6]. In addition, in the American context, because of the recency of the 2024 US Presi-

dential Elections, established public datasets are sparse. Thus, the research had to collect data through APIs and open-sourced scrapers. `X API v2` was used to retrieve relevant tweets from X (formerly Twitter) then put through a series of Python codes [5].

Ways of extracting posts from Facebook and TikTok are different from ways of extracting tweets from X (formerly Twitter) because of the unique nature of their postings, which, for example, in the case of TikTok, consist of only images or videos. In posts from Facebook, in the study from Grujic, et al. [2014], they employ the Facebook Graph API, an HTML-based API to access information from Facebook [18]. It utilized PHP, HTML, and jQuery to query data, post new stories, upload photos, and more. In the study by Alashri, et al. [2016], they used Python codes to extract Facebook posts and comments from the pages of presidential candFdates for the 2016 US Presidential Elections [3].

Meanwhile, in extracting data from TikTok, Vassey, et al. [2022] gathered data related to little cigar and cigarillo products by scraping public posts with hashtags containing high engagements [38]. Researchers then developed a codebook to analyze themes within the videos and captions, establishing inter-rater reliability through subsample coding. Similar methods are also used by Abbas, et al. [2022] wherein they used the top three hashtags (`#SaveSheikhJarrah`, `#SavePalestine`, `#FreePalestine`) to search TikTok video content related to youth

activism in the Israel-Palestine conflict, however, it also used to find videos via non-random sampling method to code frame by frame to see what is the message of the video [1]. Lastly, Cheng and Li [2023] converted the audio of news videos in TikTok to text via a Google API to train the text into a sentiment classifier [11]. Then, they took images of them every second to calculate the second-person view ratio, essential for studying the prevalence of that point of view in every news video. Then, they took every frame-per-second of a video to calculate the second-person view ratio, essential for studying the prevalence of that viewpoint.

From the existing studies mentioned, although these methods are possible, it still poses a challenge for the researchers to collect social media data due to the social media platforms' dynamic website structure, limited APIs, rate limits, and potential data noise that might be collected.

CHAPTER III

METHODOLOGY

The methodology will follow the Figure 3.1.

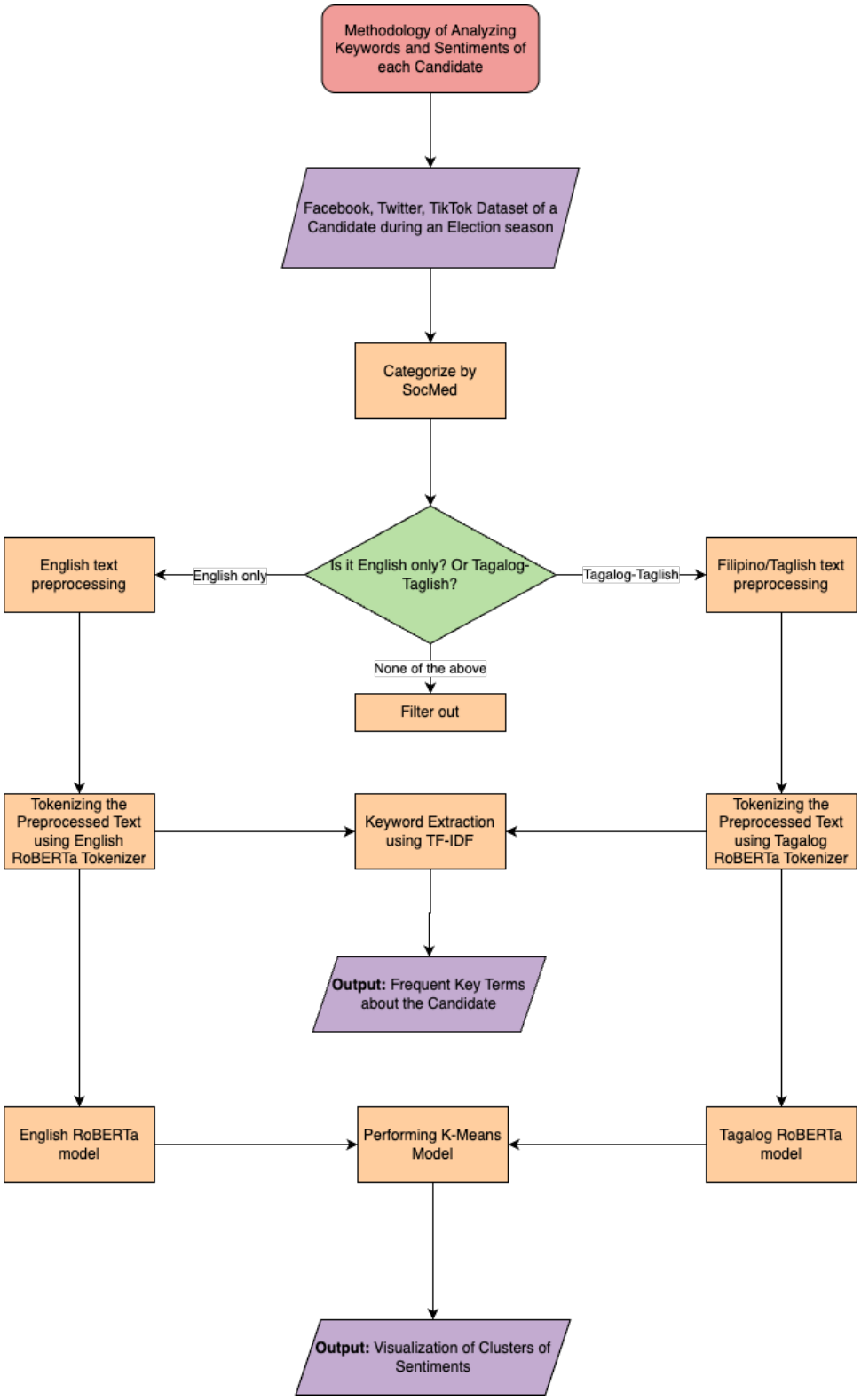

Figure 3.1: Methodology Flowchart

3.1 Data Collection

The collection of datasets will happen through the official Application Program Interfaces (APIs) of Twitter, Facebook, and TikTok. If the said APIs are unavailable or country-restricted, open-source tools will be used to collect posts and comments. Due to the unique nature of TikTok, which consists of either images or videos, it will just consist of captions of the video and comments.

The dates of the posts and comments in the data will be as follows, starting from the day one of the candidates announced their nomination up to two days before the election day. In collecting data based on the given timeframe, there should be at least one of the key terms present:

Presidential Election years	**Intention to Run Announcement**	**End Date**	**Key Terms**
2016	Duterte and Cayetano: November 21, 2015 Roxas and Robredo: July 31, 2015	May 7, 2016	Duterte, Cayetano, Mar Roxas, Roxas, Robredo, Leni, Robredo, DU30, Roro
2022	Marcos and Duterte: October 5, 2021 Robredo and Pangilinan: October 7, 2021	May 7, 2022	Marcos, Duterte, Robredo, Kiko, Pangilinan, BBM, Leni, DU30, Uniteam, Bongbong

Table 3.1: Table for Philippine Dataset Ranges

Presidential Election Year	Intention to Run Announcement	End Date	Key Terms
2020	Biden and Harris: April 25, 2019 Trump and Pence: January 21, 2017	November 6, 2020	Biden, Harris, Kamala, Trump, Pence, Joe Biden, Donald Trump
2024	Trump and Vance: November 15, 2022 Harris and Walz: July 21, 2024	November 6, 2024	Harris, Walz, Trump, Vance, Kamala, JD Vance

Table 3.2: Table for United States Dataset Ranges

Each collected post or comment mentioning a candidate(s) will be in the `.csv` file separately, with each row marked by which social media platform it comes from. They will be utilized separately according to the election year and country for text classification.

3.2 Data Preprocessing

After grouping the datasets, before they are fed into the RoBERTa tokenizer, posts and comments will undergo preprocessing before feeding the data into the RoBERTa model. Since RoBERTa recognizes the stop words such as "the," "a," "is", etc., it will not be removed during this process. The following steps to preprocess the text will be as follows:

1. Omitting texts from the dataset if it is not in English, Tagalog, or Taglish.
2. Removing punctuation marks that have no significance for sentiment analysis.
3. Replacing emojis with special tags.
4. Removing unnecessary emojis or replacing emojis with special tags describing them if it is necessary for sentiment analysis.
5. Lowercasing the text.
6. Handling links and email addresses by replacing them with a placeholder.
7. Removing whitespaces and replacing multiple spaces with a single space.
8. Adding paddings to equalize the length of sentences.

The preprocessed dataset will be placed in a `.csv` file.

3.3 Text Classification and Visualization

After the preprocessing, the dataset will be fed into the following models: the RoBERTa model for contextualized embedding and the TF-IDF model for determining frequent keywords.

Since the standard RoBERTa model primarily uses an English dataset for pretraining, to handle Tagalog and Taglish language texts, the researchers will

use the Tagalog RoBERTa model created by DOST ASTI [40]. The preprocessed dataset will be fed into the RoBERTa Tokenizer before the tokenized result will go into the embedding model. After the formulation of embeddings, they will be fed into the K-Means model so that they will be assigned to the clusters that are essential for creating the visualization of the sentiments based on the semantic similarity of the texts, visualizing the echo chambers. This will be used to compare and contrast the social media presence and activity of each candidate.

Meanwhile, the tokenized texts will also go to the TF-IDF model to determine the frequency of words. The frequency of the words is sorted by how frequently they are used in a certain post or comment. The top 30 keywords per candidate will be used to compare and contrast with other candidates, especially if they are from another country.

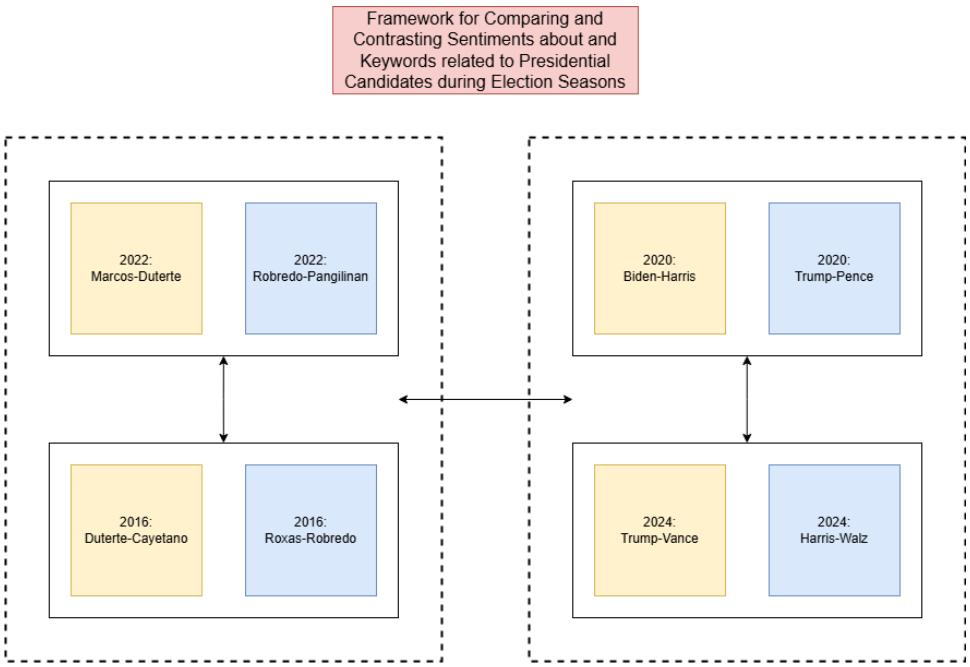

Figure 3.2: Framework for Comparing and Contrasting Sentiments

BIBLIOGRAPHY

[1] Abbas, L., Fahmy, S. S., Ayad, S., Ibrahim, M., and Ali, A. H. Tiktok intifada: Analyzing social media activism among youth. *Online Media and Global Communication 1*, 2 (2022), 287–314.

[2] ABS-CBN. Halalan results 2022. `https://halalanresults.abs-cbn.com/`, 2022.

[3] Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L., and Desouza, K. C. An analysis of sentiments on facebook during the 2016 us presidential election. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2016), IEEE, pp. 795–802.

[4] Ampon, R. K. S., and Salathong, J. 'battleground of philippine elections': Political message strategies of bongbong marcos and leni robredo on social media. *International Journal of Social Science and Human Research 6*, 07 (2023).

[5] Ancheta, J. R., Gorro, K. D., and Uy, M. A. D. Walangpasok on twitter: Natural language processing as a method for analyzing tweets on class suspen-

sions in the philippines. In *2020 12th International Conference on Knowledge and Smart Technology (KST)* (2020), pp. 103–108.

[6] Aquino, J. A., Liew, D. J., and Chang, Y.-C. Graph-aware pre-trained language model for political sentiment analysis in filipino social media. *Engineering Applications of Artificial Intelligence 146* (2025), 110317.

[7] Baccini, L., Brodeur, A., and Weymouth, S. The covid-19 pandemic and the 2020 us presidential election. *Journal of population economics 34* (2021), 739–767.

[8] BBC. Us presidential election results 2024. `https://www.bbc.com/news/election/2024/us/results`, 2024.

[9] Beaumont, T. Joe biden launches 2020 white house bid. *PBS News* (25, Apr 2019).

[10] Buan, L. Dictator's son bongbong marcos to run for president in 2022. *Rappler* (5, Oct 2021).

[11] Cheng, Z., and Li, Y. Like, comment, and share on tiktok: Exploring the effect of sentiment and second-person view on the user engagement with tiktok news videos. *Social Science Computer Review 42*, 1 (2024), 201–223.

[12] Cruz, L. C., dela Cruz, J. N., Maglangit, S. F., Magtira, M., Imperial, J. M., and Rodriguez, R. Is twitter an echo chamber? connecting online public sentiments to actual results from the 2019 philippine midterm elections. In *2022 International Conference on Asian Language Processing (IALP)* (2022), IEEE, pp. 57–62.

[13] Cupin, B. Mar roxas launches 2016 presidential bid. *Rappler* (31, Jul 2015).

[14] Demillo, R. E., Solano, G., and Oco, N. Philippine national elections 2022: Voter preferences and topics of discussion on twitter. In *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (2023), pp. 724–729.

[15] Dong, X., Li, C., and Choi, J. D. Transformer-based context-aware sarcasm detection in conversation threads from social media, 2020.

[16] Dulay, D., Hicken, A., Menon, A., and Holmes, R. Continuity, history, and identity: Why bongbong marcos won the 2022 philippine presidential election. *Pacific Affairs 96*, 1 (2023), 85–104.

[17] Gold, M. President trump tells the fec he qualifies as a candidate for 2020. *The Washington Post* (21, Jan 2017).

[18] Grujic, I., Bogdanovic-Dinic, S., and Stoimenov, L. Collecting and analyzing data from e-government facebook pages. *ICT Innovations* (2014), 86–96.

[19] Holmes, R. D. The dark side of electoralism: Opinion polls and voting in the 2016 philippine presidential election. *Journal of Current Southeast Asian Affairs 35*, 3 (2016), 15–38.

[20] International, A. Five things to know about martial law in the philippines. *Amnesty International* (21 Sep 2022).

[21] Jazeera, A. Donald trump to face 'hush money' criminal trial: What's it all about? *Al Jazeera* (15 February 2024).

[22] Johnson, H., and Head, J. Leni robredo: The woman leading the philippines' 'pink revolution'. *BBC* (7 May 2022).

[23] Koroteev, M. V. Bert: A review of applications in natural language processing and understanding, 2021.

[24] Lalu, G. P. 'buong-buo ang loob ko': Robredo to run for president in 2022. *Inquirer.Net* (7, Oct 2021).

[25] Lee, C. Kamala harris is using social media to reach young voters. *TIME* (3 Sep 2024).

[26] LIU, B. *Sentiment analysis and opinion mining*. SPRINGER INTERNATIONAL PU, 2012.

[27] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019.

[28] Macrohon, J. J. E., Villavicencio, C. N., Inbaraj, X. A., and Jeng, J.-H. A semi-supervised approach to sentiment analysis of tweets during the 2022 philippine presidential election. *Information 13*, 10 (2022).

[29] Mendoza, M. E. H. Philippine elections 2022: Tiktok in bongbong marcos' presidential campaign. *Contemporary Southeast Asia: A Journal of International and Strategic Affairs 44*, 3 (2022), 389–395.

[30] Orr, G., Holmes, K., and Stracqualursi, V. Former president donald trump announces a white house bid for 2024. *CNN* (16, Nov 2022).

[31] Pernia, R. A., and Panao, R. A. L. Electing the dictator's son: the 2022 philippine election in an era of authoritarian nostalgia and democratic decline. *Asian Affairs: An American Review* (2025), 1–22.

[32] Pranay Kumar, B., and Sadanandam, M. A fusion architecture of bert and roberta for enhanced performance of sentiment analysis of social media plat-

forms. *Manchala, A Fusion Architecture of BERT and RoBERTa for Enhanced Performance of Sentiment Analysis of Social Media Platforms* (27, May 2023).

[33] Prasanthi, N., Madhavi, R., Sabarinadh, D., and Sravani, B. A novel approach for sentiment analysis on social media using bert and roberta transformer-based models. pp. 1–6.

[34] Ranada, P. Rodrigo duterte: I am running for president. *Rappler* (21, Nov 2015).

[35] Setiawan, D., Ananda, D., and Kartika, T. Media framing of donald trump's 2024 election victory: A case study on international media. *MEDIASI Jurnal Kajian dan Terapan Media, Bahasa, Komunikasi 6*, 1 (2025).

[36] Sinpeng, A., Gueorguiev, D., and Arugay, A. A. Strong fans, weak campaigns: Social media and duterte in the 2016 philippine election. *Journal of East Asian Studies 20*, 3 (2020), 353–374.

[37] Suciu, P. Social media proved crucial for joe biden – it allowed him to connect with young voters and avoid his infamous gaffes. *Forbes* (17 Nov 2020).

[38] Vassey, J., Donaldson, S. I., Dormanesh, A., and Allem, J.-P. Themes in tiktok videos featuring little cigars and cigarillos: Content analysis. *Journal of Medical Internet Research 24*, 11 (2022).

[39] Viner, K. Read kamala harris's full statement: 'my intention is to earn and win this nomination'. *The Guardian* (21, Jul 2024).

[40] Visperas, M. L., Borjal, C. J., Adoptante, A. J. M., Abacial, D. S. R., Decano, M. M., and Peramo, E. C. iTANONG-DS : A collection of benchmark datasets for downstream natural language processing tasks on select Philippine languages. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)* (Online, Dec. 2023), M. Abbas and A. A. Freihat, Eds., Association for Computational Linguistics, pp. 316–323.