# Research paper - Election Analysis and Winner Prediction Via Twitter

Tianhao Xu[2740825], Lu Wang[2754874], Zheng Zhang[2748843], and Han Lin[2724786]

Vrije Universiteit Amsterdam
The Social Web
Group 18

**Abstract.** Elections are a crucial aspect of every citizen's life and the primary democratic tool. The institutions or entities (political parties or particular politicians) that should represent the people are chosen by the people. Pre-election surveys have historically been used to identify trends and forecast election outcomes. It has become quite popular to forecast election outcomes using user activity on the social networking site Twitter. But earlier research has concentrated on predicting elections in developed countries like the United States, the United Kingdom, Spain, and France. The article specifically looks at how elections were run in developing countries using Kenya as its research subject. This study looks at relevant tweets from Kenya that were collected between June 2022 and July 2022. This paper addressed three research questions that mainly used unsupervised learning in machine learning and sentiment analysis. According to the final results, William Ruto was predicted to win the election.

**Keywords:** Sentiment analysis · Twitter · Presidential election · Prediction · Machine learning

## 1 Introduction

Pre-election polls are a crucial component of elections because they give voters and political organizations the chance to modify their voting or campaign strategy as needed. Since the early 19th century, opinion polls have been conducted over. As prediction models have developed over time, they are now capable of accurately forecasting elections. Election polling failures happen frequently in developing nations like Kenya, although they are comparatively uncommon in industrialized nations [6].

Traditional polls are now very expensive. The use of social networking sites as platforms for discussing political opinions is growing, and researchers are beginning to use social networking sites as viable sources for poll mining. In this case, user opinions are inferred from the material and the connections they establish over time rather than being expressly requested in a set of predetermined questions. Twitter in particular has developed to become a source of election prediction for a variety of surveys due to the data availability [7].

The primary goal of this article is to examine the following three research issues using unsupervised machine learning and sentiment analysis to predict the Kenyan elections in 2022:

- What are the keywords surrounding the presidential candidates?
- Which candidate can win through analysis?
- What kind of sentiment do people have about the candidates?

We discovered that citizens desire to know more about the candidates they support through collecting and analysing feedback on the general election in Kenya. Our study intends to understand the true intentions of the candidates, the public's expectations of the candidates, and predict the winner with the margin of victory. It hoped that we could help countries to regulate the election from the perspective of social media discourse. This will effectively ensure that further elections are fair and reliable.

The article is organized as follows: Section 2 reviews the literature on traditional election forecasting models, election prediction based on Twitter sentiments, and the comparison between Twitter sentiments and poll data. Section 3 discusses the method used to answer the research question. Section 4 introduces the data used in this study, discusses data preprocessing techniques, and uses data visualizations to do the data analysis. Section 5 interprets the results. Finally, this paper concludes with a discussion of findings, privacy & ethics, Data security, and research limitations in section 6.

## 2 Literature

We briefly studied the literature in these related fields for the sake of this study, including election prediction models in politics, election prediction based on Twitter sentiment, and a comparison of Twitter sentiment with polling data.

### 2.1 Election prediction models in politics

Lewis-Beck and Rice [8] created the first presidential election prediction model. The gross national product (GNP) growth rate and the president's job approval rating were both taken into account by the model as potential predictors. The trialheat model was created by Campbell and Wink [9] using this model as a foundation. The support for the candidate of the ruling party and the growth rate of real GDP was also included in the model as predictor variables. The group created the convention-bump model in 1992, which takes into account three predictors, including support for candidates in candidate polls, changes in party candidate support, and GDP growth rates. [10]

### 2.2 Election prediction based on Twitter sentiments

When analyzing the sentiment of election-related Twitter data, there are two main methods. One method uses lexicons, and the other uses machine learning.

Dictionary-based sentiment analysis techniques compare the frequency or occurrence of terms in a text with words in the dictionary using a predefined sentiment dictionary. As an illustration, Ahmed, Jaidka, and Skoric projected elections in four nations and contrasted the accuracy of their predictions with the significance of various technological infrastructures and democratic processes in each nation. [14] They used SentiStrength, a sentiment dictionary, to assign positive and negative scores to all tweets related to a party for sentiment analysis.

Unsupervised learning techniques and supervised learning techniques are the two main kinds of machine learning techniques used for sentiment analysis. The majority of previous research has utilized supervised learning techniques, which call for thoroughly pre-labeled training datasets. For the 2012 US presidential election, Wang et al. created a system for real-time analysis of Twitter sentiment toward presidential candidates. [12] To categorize sentiment, they trained a straightforward Bayesian model on unigram features. In the six months leading up to the 2016 US presidential election, Paul et al. gathered geo-tagged tweets and categorized them as Democratic or Republican based on their attitude at the county level. [13] On 1.6 million tweets from the Stanford Twitter Sentiment (STS) dataset, they trained Support Vector Machines (SVM), Polynomial Parsimonious Bayes (MNB), Recurrent Neural Networks (RNN), and FastText models.

### 2.3 Twitter sentiments and poll data comparison

Previous studies have demonstrated the feasibility of using Twitter sentiment in place of opinion polls. O'Connor et al. [16] analyze numerous consumer confidence and political opinion polls from 2008 to 2009 and discover a correlation between Twitter sentiment and general public opinion. During the 2012 presidential election, Beauchamp [15] modelled state-level polls as a function of political tweets and discovered that Twitter-based indicators predicted opinion surveys. Election polls based on Twitter can give more accurate results than conventional polls, particularly in developing nations, as demonstrated by Nugroho et al. [11]

## 3 Method

This section focuses on the project pipeline and the methodologies we employed to investigate the responses to our research questions.

### 3.1 Project Pipeline

The two key components of our project are data extraction and processing, followed by sentiment analysis and election prediction. As seen in Figure 1, the data will be pre-processed in the first part, and the processed data will be divided into several sub-datasets as per candidate; the specifics of the data pre-processing will be covered in more detail in Section 4. In the second part, the sentiment analysis uses processed data whose results reflect public sentiment toward the

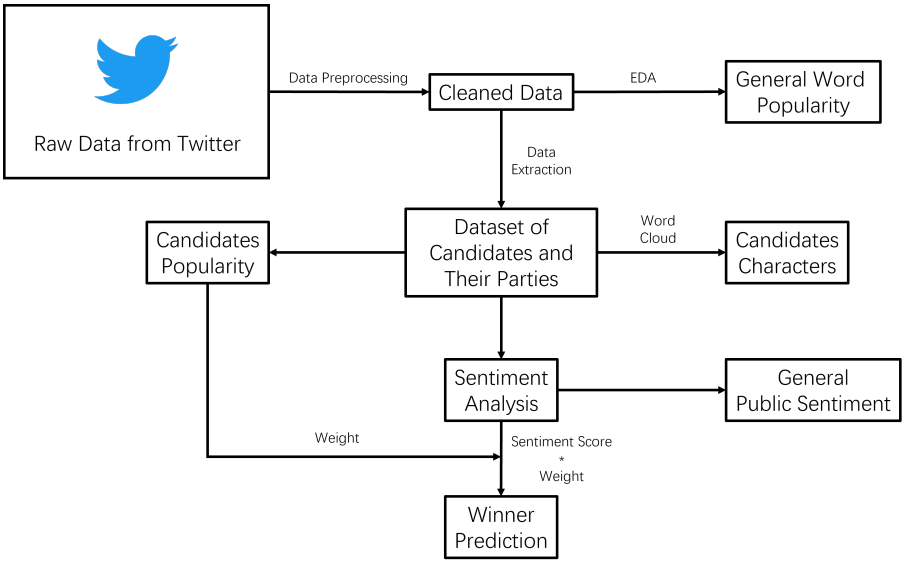


**Fig. 1.** Project pipeline

candidates. The results of the sentiment analysis together with public interest determine how the general election will turn out.

### 3.2 Data Extraction and Candidates Analysis

Some names in the dataset need to be replaced before the data can be extracted and segmented by candidates. This is done for two reasons, firstly candidates sometimes go by several identities, like "deputy" and "ruto," all of which refer to "William Ruto." Therefore, the official name is used in place of these "nicknames." Second, the same circumstance occurs when referring to a party's name. For instance, the party "Kenya Kwanza" is referred to by both "uda" and "hustler nation." Then those relevant keywords will also be replaced by the official party names. In order to improve the effectiveness of the program and decrease data noise, remove all records that do not contain the name of the relevant candidate or political party. We can then count the number and percentage of records associated with each candidate and party, and the percentage of records in which each candidate is mentioned is considered to be the public profile of that candidate. The dataset is then divided into a number of smaller datasets according to the names of the candidates. The word cloud is applied to generate characteristic words-cloud of each candidate, which is a kind of weighted list to visualize language or text data [17].

### 3.3 Sentiment Analysis and Winning Prediction

TextBlob [18], a library in Python 2 and Python 3, is the model for sentiment analysis and is employed to process textual data. This library provides an easy-to-use API for exploring Natural Language Processing (NLP) topics. The main benefit of using TextBlob is that it is built on both NLTK and pattern and works well with both. It also offers a variety of features, including stage extraction, post-tagging, estimation testing, etc. Alongside the deployment of Textblot, a new feature named polarity is generated, which lies between [-1,1]. Positive emotions are represented by a value larger than 0, whereas negative emotions are represented by a value less than 0, and neutral emotions are represented by a value equal to 0.

We will tally the number of records for each candidate's sub-dataset of the three emotions, and then assign each candidate a score based on public sentiment and attention. The equation of the scoring algorithm is:

$$Score = \mathbf{NumberofPostive} + 0.5 * \mathbf{NumberofNeutral} - \mathbf{NumberofNegtive}.$$

## 4 Data

In this section, we describe the dataset used for the project and the pre-processing of the data prior to conducting the sentiment analysis.

### 4.1 Dataset Overview

About 400,000 tweets, each containing a tweet id, tweet time, tweet text, and retweet count, make up the dataset utilized for this project, which was released in July 2022, and it gathered tweets from June to July 2022. The accuracy of the data in the selected dataset will be assumed.

### 4.2 Data Pre-processing

When we viewed the dataset, we noticed that there were a significant number of empty cells; those required to either be filled in with alternative values or removed. About 26.69% of the values for tweet id, time, and retweet count are missing. As tweet id is a unique identifier that should be generated by Twitter, it will disregard since we are unable to infer them. Backfill was used to fill up every missing value in column time. The record with missing value in column tweet will be removed because other attributes will be meaningless once there is no content to analyze. The missing values in retweet_count were replaced with 0. Since the percentage of missing values in the data frame is less than 0.01%, then this record was removed. In order to make the tweets easier to analyze, we process the tweets by removing special characters, double spaces, and duplex tweets, converting the text's case to lowercase, and removing unneeded links and photographs. After pre-processing, there has been a loss of 12.87% of all records, leaving 353,439 records in the data frame. The processed data were tokenized and lemmatized to make it easier for conducting sentiment analysis.


**Fig. 2.** General word popularity

### 4.3 Exploratory Data Analysis

After extracting clean data, we carried out some data exploration with the primary goal of understanding the fundamental viewpoints on the election discussion, which we attained by generating word clouds and keyword statistics as shown in Figure 2 and Figure 3.

## 5 Result

First, from the word cloud of all election-related tweets (see Figure 2), we found that in addition to the candidates' names, there were other keywords that appeared with high frequency, such as "th" (ordinal word), "root party", "grade", etc. To better analyze the reasons behind this and to find out the keywords surrounding the presidential candidates, we looked at the more detailed word cloud around each election candidate separately (see Figure 4).

According to the results of the study, among the keywords associated with candidate David Waihiga Mwaure, "azimio", "economist", "academic" and "debate" were the most frequently used words, indicating that people's attitudes towards Waihiga Mwaure were not particularly much evaluated in terms of emotional tendencies, but reflected his personal achievements and indicated that his party The most frequently used keywords associated with candidate George Wajackoyah were 'azimio', 'clever', 'stop', 'go', 'hate' etc. This suggests that people have both positive and negative views of Wajackoyah, and also reflects the fact that he is associated with the azimio party. For Raila Odinga, the keywords often

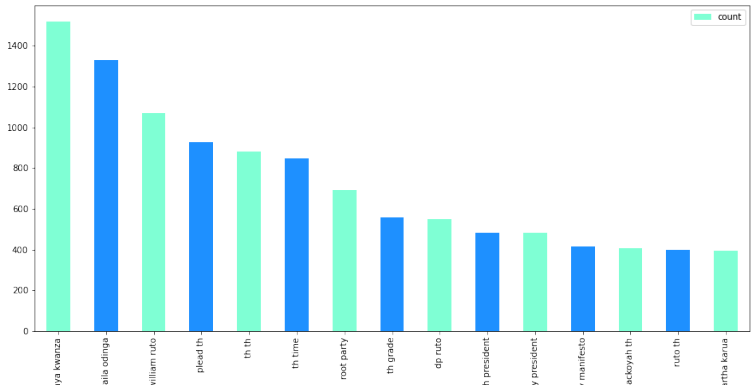


**Fig. 3.** Distribution of Top 15 Words

used were "baba", "azimio", "politics", "tumbocrats", "kesho" (meaning tomorrow in English) and "continue", indicating that people had many more positive emotions to comment on him than negative ones. Finally, in the word cloud for William Ruto, the words "local", "prime", "thank", "politics" and "shame" appear to occur frequently, suggesting that people may have a high opinion of Ruto, with positive words appearing more often.

The words political party and sentiment appear several times in the word cloud, so we expect a more exact analysis of people's attitudes towards the candidates, so we used TextBlob to perform a sentiment analysis of the candidates. By comparing the text corpus, we divided the content of tweets comments in the dataset into three categories: positive, negative, and neutral.

By looking at the overall sentiment distribution (as shown in Figure 5), more than half of the comments were neutral, while among the comments with a sentiment tendency, the number of positive comments was twice as high as the number of negative comments, indicating that people's attitudes towards the candidates remained generally supportive, although there was still a significant proportion of people with negative attitudes towards the candidates.

Therefore, we showed people's sentiments towards each candidate according to the different candidates which can answer our research question of "What kind of sentiment do people have about the candidates?" (see Figure 6). We found that this result is similar to the results of the related word cloud analysis discussed earlier, there were very few sentiment ratings about Waihiga Mwaure compared to the other three candidates, and it can be deduced that people pay very little attention to him. In Figure 6, we can see that both candidates Ruto and Odinga have a high level of popularity, with a similar number of positive and negative sentiments directed at them, and people are generally more supportive than opposed to them. For candidate Wajackoyah, the number of positive and

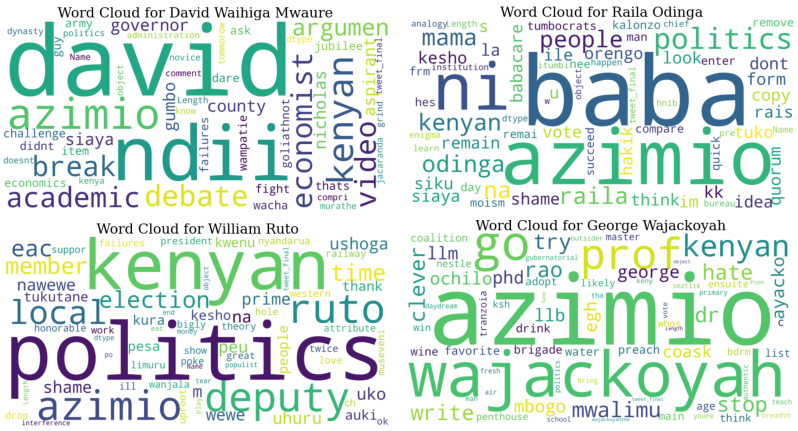


**Fig. 4.** Word Cloud for each Candidate

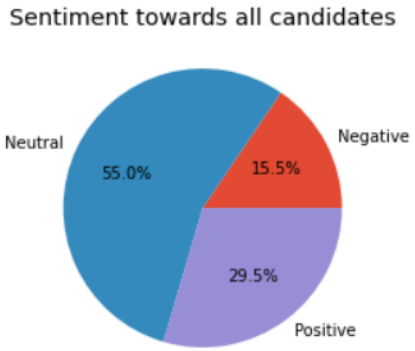


**Fig. 5.** Total Sentiment

negative comments is almost equal, and it is clear that people's opinion of him is mixed.

Finally, to answer our research question of "which candidate will win in the election", we calculate each candidate's score based on the sentiment label of each candidate and use this to predict the candidate most likely to succeed in the presidential election. In the calculation, we gave 2 points to positive sentiment comments, -1 point to negative sentiment comments, and 1 point to neutral comments (although neutral comments should not have an impact on the score, the neutral comments here are comments about a specific candidate and therefore reflect the level of interest in a candidate, so neutral comments should also be given a positive value to influence the score). The scores of the three candidates are shown in Table 1, and we can see that Ruto received the highest score, so our prediction is that Ruto will be elected president.

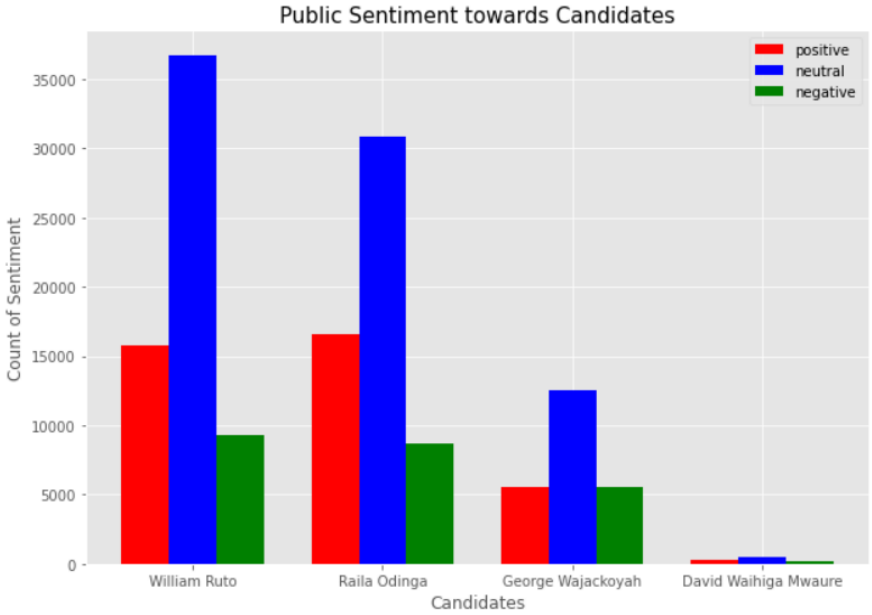


**Fig. 6.** Sentiment towards Candidates

## 6 Conclusion

This section talks about the summary of the main findings of the election in Kenya in 2022. We also demonstrate the problem of privacy, data security, and ethics with mitigation measures. Then the novelty of this article are been provided. In the last part, the limitations of our research are been illustrated.

### 6.1 Summary of the main findings

Based on the analysis of results we presented earlier, we can find that although Ruto and Raila were close in the number of positive and negative sentiment

| Candidate's Name | Score |
|---|---|
| William Ruto | 43198 |
| Raila Odinga | 38785 |
| George Wajackoyah | 15757 |
| David Waihiga Mwaure | 646 |

**Table 1.** Candidates' Prediction

comments, and even though Raila had a higher number of positive comments and a lower number of negative comments, the final predicted outcome was a win for William Ruto. We found that this was largely dependent on the higher number of neutral sentiment comments around Ruto. It is not difficult to conclude that the prediction of Ruto's election to the president was not only based on the percentage of positive and negative comments surrounding him, but also on the number of neutral comments. This is because a more neutral attitude toward comments around them means that they receive more attention, which indirectly affects his social presence, and this is why Ruto's won by a small margin.

According to the recent news of the election result, it was the same result as we predicted, Ruto won with 50.49% of the vote, narrowly defeating Raila, who got 48% of the vote. So it proves that our simulated result is a valid prediction.

### 6.2 Novelty

Our study was picked for a developing country that Kenya with new dataset even though the majority of election-predicting projects are targeted toward developed countries. For data processing, in order to preserve as much relevant information as possible in the data, we replaced the nicknames of many candidates with their official names during the data processing procedure. In addition to considering public sentiment, our scoring also factors in how much the public is paying attention to the predictions, which has been shown to increase prediction accuracy.

### 6.3 Privacy and Data Security

The data collection stage of our research presents the most privacy risk. Numerous privacy concerns may arise during the data collecting and processing process. Data can be used by those with bad intentions to harm society in some ways. We need to gather a lot of data for our academic purposes in order to perform our study, but user privacy and data security must come first. Our research questions include changes in people's sentiment towards candidates, the most commonly used keywords, and the proportion of fake news present during the election process, among others. To address these questions, we need to collect a large number of posts, analyze their data and evaluate their content. For us, though, what matters is the content of the statements, not who is uttering them. For this reason, we decided against using any post quotes in our articles. to safeguard the privacy of the users whose information we gather for our study. Additionally, this demonstrates respect for each citizen's political beliefs and inclinations. We do not want to see political disputes among individuals if this occurs.

There are numerous laws and regulations that control the gathering and keeping of personal data. The General Data Protection Regulation (GDPR) is the most well-known EU statute governing privacy laws. [3] The content of public posts and the posting date are the only pieces of data we gather to fulfill

our study goals. No additional data is gathered. This data is exempt from the stringent GDPR laws because it is not considered personal data.

We gather information via the Twitter API to decrease privacy threats. Twitter is not permitted to share any personal information with us because they must abide by GDPR. We only see anonymized data. In actuality, we are unsure of who posted it. This indicates that we never acquire personal information without consent. We review all the results that come up to make sure we are in compliance and are not gathering any data that is not permitted.

### 6.4 Ethics

There are a wide variety of potential defences for elections. Personal lives are discussed, along with political opinions, religious convictions, morals, and values. People can talk politics and share their opinions of the candidates on Twitter. When discussing elections, we should think about whether it is moral to disregard the justifications and opinions of others. Additionally, there is fake news on social media, and people interfere in elections and spread false information on Twitter. It becomes important how we identify them. Using false information to carry out our research is unethical.

To prevent these circumstances, we must take some action. As an illustration, gaining ethical approval and anonymizing data were permitted. They are crucial to preventing potential ethical issues during the research procedure. Researchers must examine each post on Twitter in order to study the electoral discourse there. This implies that a large number of people may be impacted, and their privacy may be compromised because their names and other personal information may be revealed in the analysis report. These folks might not be aware that their posts are being looked into, and they might not agree. Furthermore, discussions on Twitter platforms can start contentious arguments regarding elections, which can lead to a variety of favorable or unfavorable outcomes. Therefore, it's critical to comprehend the research topic's ethical implications and the potential negative effects of any privacy violations.

### 6.5 Limitations of our research

Our work on Twitter-based election prediction has some limitations. We discovered that the accuracy of predictions was significantly impacted by data preprocessing and data selection. The majority of the experiments were carried out after the election, and the outcomes served as the underlying data for the predictions. The results' generalizability is so constrained. In addition, social networks themselves suffer from bias, concerns with credibility, and other problems that call for great attention when making election predictions.

Election polls conducted via Twitter have a higher degree of accuracy than traditional polls, especially in developing countries. [1] In this case, we must concentrate on de-biasing the information that was collected. [2] Demographic bias is much higher among Twitter users (young, male, and living in urban areas) in developing countries compared to developed countries. For example, in our

crawling exercise, we collected over 30,000 tweets from users living in the capital city of Nairobi (4.4 million inhabitants), whereas we were only able to collect 50 tweets from users living in Kakamega (over 1.8 million inhabitants).

We will eliminate bias in the future work. For example, we could eliminate data bias based on user age by crosswalking full names and counties of residence with online public records. Alternatively, we could take into account other data sources. Improve Twitter-based predictions by considering the proportion of the population and social media users in each age group as well as each candidate's vote share in the last general election.

## References

1. Dwi Prasetyo N, Hauff C. Twitter-based election prediction in the developing world[C]//Proceedings of the 26th ACM Conference on Hypertext  Social Media. 2015: 149-158.
2. M. Choy, M. L. Cheong, M. N. Laik, and K. P. Shung. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. arXiv preprint arXiv:1108.5520, 2011.
3. Voigt P, Von dem Bussche A. The eu general data protection regulation (gdpr)[J]. A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017, 10(3152676): 10-5555.
4. D. S. Hillygus. The evolution of election polling in the United States. Public opinion quarterly, 75(5):962–981, 2011.
5. M. S. Lewis-Beck. Election forecasting: principles and practice. The British Journal of Politics  International Relations, 7(2):145–164, 2005.
6. L. Fumagalli and E. Sala. The total survey error paradigm and pre-election polls: The case of the 2006 Italian general elections. Technical report, Iser Working paper Series (No. 2011-29), 2011.
7. M. Choy, M. L. Cheong, M. N. Laik, and K. P. Shung. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. arXiv preprint arXiv:1108.5520, 2011.
8. Lewis-Beck, M. S., and T. W. Rice. 1982. "Presidential Popularity and Presidential Vote." Public Opinion Quarterly 46 (4): 534–537. doi:10.1086/268750.
9. Campbell, J. E., and K. A. Wink. 1990. "Trial-Heat Forecasts of the Presidential Vote." American Politics Quarterly 18 (3): 251–269. doi:10.1177/1532673X9001800301.
10. Campbell, J. E., L. L. Cherry, and K. A. Wink. 1992. "The Convention Bump." American Politics Quarterly 20 (3): 287–307. doi:10.1177/1532673X9202000302.
11. Liu R, Yao X, Guo C, et al. Can we forecast presidential election using twitter data? an integrative modelling approach[J]. Annals of GIS, 2021, 27(1): 43-56.
12. Wang, H., D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. 2012. "A System for Real-time Twitter Sentiment Analysis of 2012 Us Presidential Election Cycle." In Proceedings of the ACL 2012 System Demonstrations, Jeju Island, Korea, 115–120.
13. Paul, D., F. Li, M. K. Teja, X. Yu, and R. Frost. 2017. "Compass: Spatio Temporal Sentiment Analysis of US Election What Twitter Says!" In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1585–1594. doi:10.1145/3097983.3098053.

14. Ahmed, S., K. Jaidka, and M. M. Skoric. 2016. "Tweets and Votes: A Four-country Comparison of Volumetric and Sentiment Analysis Approaches." In Tenth International AAAI Conference on Web and Social Media, Cologne, Germany.
15. Beauchamp, N. 2017. "Predicting and Interpolating State-level Polls Using Twitter Textual Data." American Journal of Political Science 61 (2): 490–503.doi:10.1111/ajps.12274.
16. O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." Icwsm 11 (122–129): 1–2.
17. Jin Y. Development of word cloud generator software based on python[J]. Procedia engineering, 2017, 174: 788-792.
18. Octaria O, Manongga D, Iriani A, et al. Mining Opinion Based on Tweets about Student Exchange with Tweepy and TextBlob[C]//2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). IEEE, 2022: 102-106.

# 7 Appendix

## 7.1 Contribution

Our group cooperation was pleasant, we had a full discussion and cooperation. Through group cooperation, we made up for each other's shortcomings and achieved results that we were very satisfied with. Each team member is very active, and everyone is making contributions to the homework as much as possible. For each assignment (A1, A2, A3, A4), we all fully discussed and studied. A1 is mainly completed by Tianhao Xu, A2 is mainly completed by Lu Wang, A3 is mainly completed by Han Lin, and A4 is mainly completed by Zheng Zhang.

In the peer review part, we fully read the paper and conducted in-depth discussions, and completed the relevant reports equally.

In the final report, each of us put forward several related topics, discussed and analyzed these topics, and finally determined our plan. In detail, Tianhao Xu first proposed two topics. The first topic is the authenticity of news in the Spanish corpus. The second idea is about the extraction of the relationship between characters in superheroes. Later, Zheng Zhang and Lu Wang proposed another topic about news, that is, the situation of the presidential election in the news and the recent Kenya presidential election can be our goal. Later, all of us had a team meeting and detailed our ideas, including public sentiment analysis and presidential prediction, collected relevant Twitter data and preprocessed the text. Lu Wang started our first jupyter notebook and read Twitter as the input. Tianhao Xu and Lu Wang analyzed the data, and then Zheng Zhang and Lin Han visualized the data. In this process, we discussed our data together and constantly revised our plan. Later, Lu Wang investigated Kenya's presidential candidates, got their nicknames, and programmed the generation of presidential candidates. After that, Tianhao Xu determined the political parties of these candidates and programmed them. Lu Wang and Tianhao Xu linked the candidates to their political parties. Zheng Zhang and Han Lin extracted and processed the texts of these presidential candidates, and determined that

the presidential candidates in that line had the highest number of mentions. And the data statistics of these situations are carried out. Lu Wang investigated and determined the algorithms of emotion analysis, Tianhao Xu implemented these algorithms, Zheng Zhang analyzed the data results of emotion analysis, and together with Han Lin determined the judgment conditions, and carried out a visual analysis of the results of emotion analysis. Zheng Zhang also designed a scoring algorithm that uses the results of sentiment analysis and voter enthusiasm and uses the scores to make campaign predictions for each candidate. Lu Wang and Tianhao Xu extracted and processed the relevant texts of each presidential candidate, obtained the popular words of each candidate, and discussed them with all members of the team.

We analyzed our experimental results together, modified some technology selection methods, and jointly modified the code to make them more concise. We conducted programming and report preparation on an equal footing, and our cooperation was very happy. We have learned a lot from this course. Many thanks to Professor Davide Ceolin and TAs.