ELECTION ANALYSIS USING MACHINE LEARNING

Brendon Achoki

Supervisors

Nikita Njoroge

William Okomba

Data Science

Professional Development

1.0 BUSINESS UNDERSTANDING

Business Overview

Elections are a major part in every citizen's life. People participate in choosing the leaders in their countries. Elections are an interesting thing to analyze since many people would like to know about matters surrounding their favourite candidates. Investors such as politicians may also be interested in knowing the trends and insights from such a project.

This project covers two aspects. First, to analyze the qualities that can be associated with a good leader in Kenya. An analysis of through people's sentiments was done to clarify what are the most popular qualities that the public would wish to see in their leaders are and what they want most from them. This is because politicians mention many different things that they would want to do for the citizens. Some mention good roads, reduction of prices of commodities, more schools to be built, healthcare systems among others.

The second aspect is to analyze which presidential candidate will win and by what margin. This part will not be for public consumption since it is very sensitive. It is intended that this part of the project will be handed to a company such as Ipsos so that they can integrate it in their analysis. Access to this information will be restricted to a few people.
In this project, data was scrapped from Twitter.

Problem statement

In Kenya today, presidential candidates campaign and mention the things that they will do for Kenyans once they are elected in. Also, with the tight contest in the presidential candidates, it is difficult to know who the winner will be. Therefore , this preject seeks to show what exactly it is that Kenyans want from their leaders as well as predicting the winning candidate and by what margin they win.

Objectives

General objective

- Election analysis using unsupervised learning

Specific objectives

- To analyze, through people's sentiments, what are the key things that the people wish most from their leaders.
- To predict which presidential candidate wins the election and by what margin. And also to determine if the candidate will win in the first round.

Research questions

1. What are the major issues the electorates want to be addressed by presidential candidates?
2. Which candidate is likely to win the Presidential election?

Resource inventory

The major resource needed is data which is being scrapped from Twitter. [link to dataset]

Software used

1. Google Colaboratory Notebooks.
    - Presidential prediction
    - Election analysis
    - Clustering

Constraints

Difficulty in finding a dataset for training the model.

Project plan

We will use the Cross Industry Standard Process for Data Mining (CRISP - DM) for conducting the research. We will also use the Jira Kanban board to manage the different tasks involved in this project.

2.0 DATA UNDERSTANDING

In this section, the data was explored. Data was scrapped from Twitter and made into a dataset. It had information of tweets of mostly kenyans on their sentiments concerning the 2022 presidential elections dating from January 2021.

Data description

Verifying Data Quality

The following criteria was used to clarify the quality of our data

- **Accuracy:**. The data is accurate
- **Relevancy:** It meets the requirements for its purpose
- **Completeness:** It does not contain any missing observations
- **Timeliness:** The data is up-to-date.
- **Consistency:** The format of the data is consistent

Tabular data details for the dataset

Total number of observations	405650
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	76 MB

3.0 DATA PREPARATION

Loading of data

The data was saved in a csv file, which was used when needed to load the data.
The datasets were read in and concatenated into one dataframe of 342,947 observations

Data Cleaning

Missing values and duplicate values were checked for. Missing values were imputed using backfill for time values and where retweets were missing, a zero value was placed instead.

Where there were missing values for tweets and tweet id, the observations were dropped since they would not be imputed. Unnecessary columns and 25,483 duplicated tweets were dropped. After cleaning, the data frame had 312,819 observations and 4 columns.

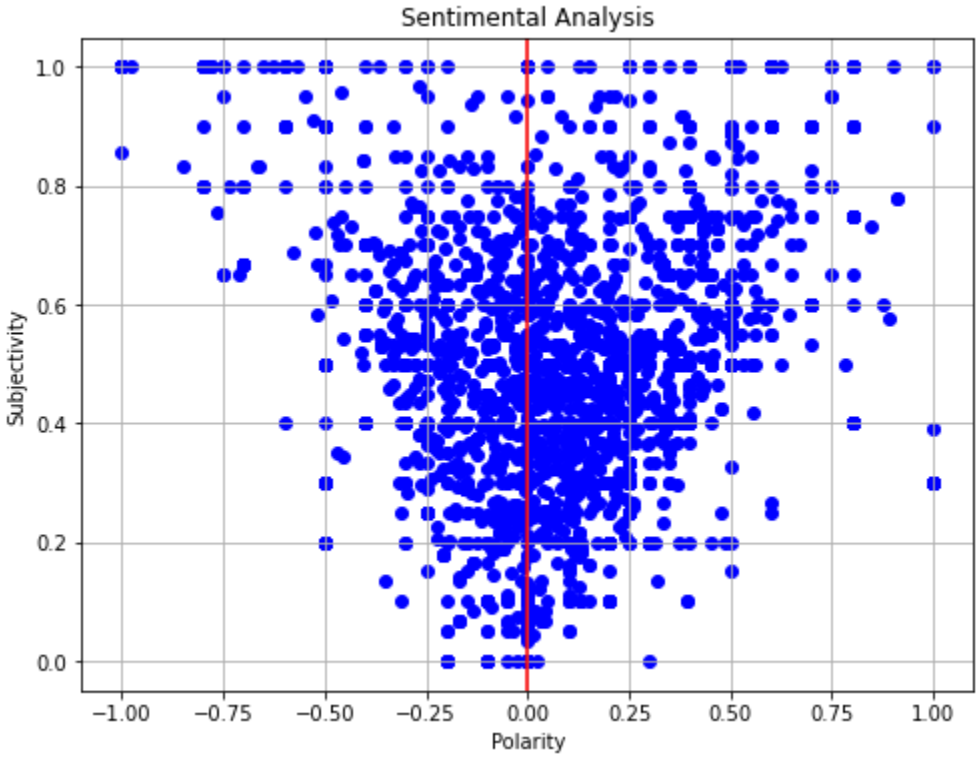
Values of Missing Values

Columns	**Values**	**Values(%)**
Tweet_id	6545	1.91
Time	157677	45.98
Tweet	29	0.01
Retweer_count	157677	45.98
TOTAL	**321,928**	**93.87**

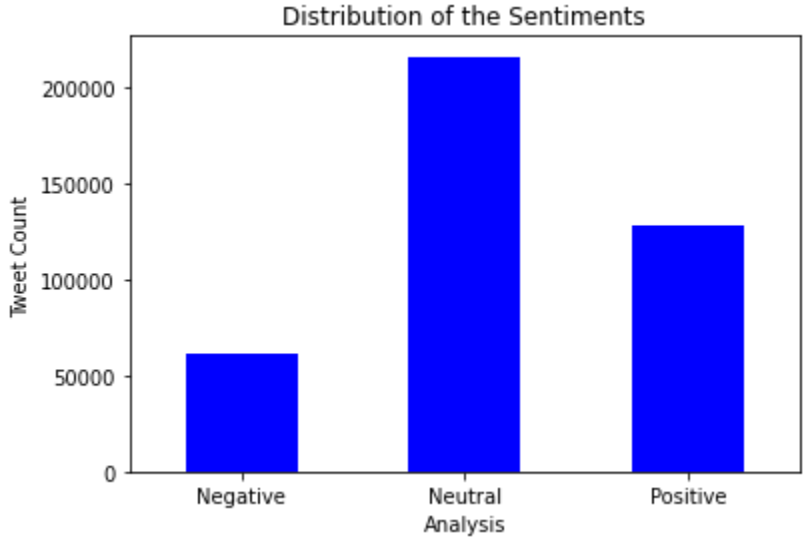
4.0 DATA ANALYSIS

Sentiment analysis

The subjectivity and polarity of tweets was checked and the result was loaded into a datafame that was used for their plotting (polarity vs subjectivity).
Positive, neutral and negative sentiments were computed from the data.



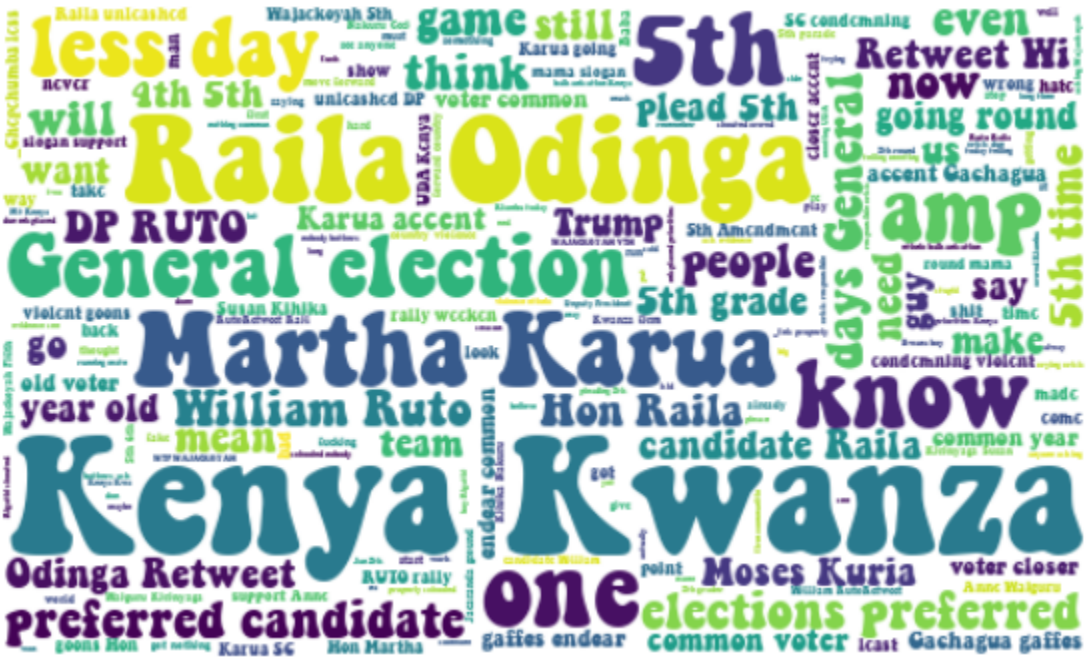
34% of the tweets were positive, 15.2% were negative, while 50.9% were neutral tweets. Below is the distribution graph



The following word clouds were plotted to show the words used in the positive, negative and neutral sentiments.

Most words used in positive tweets

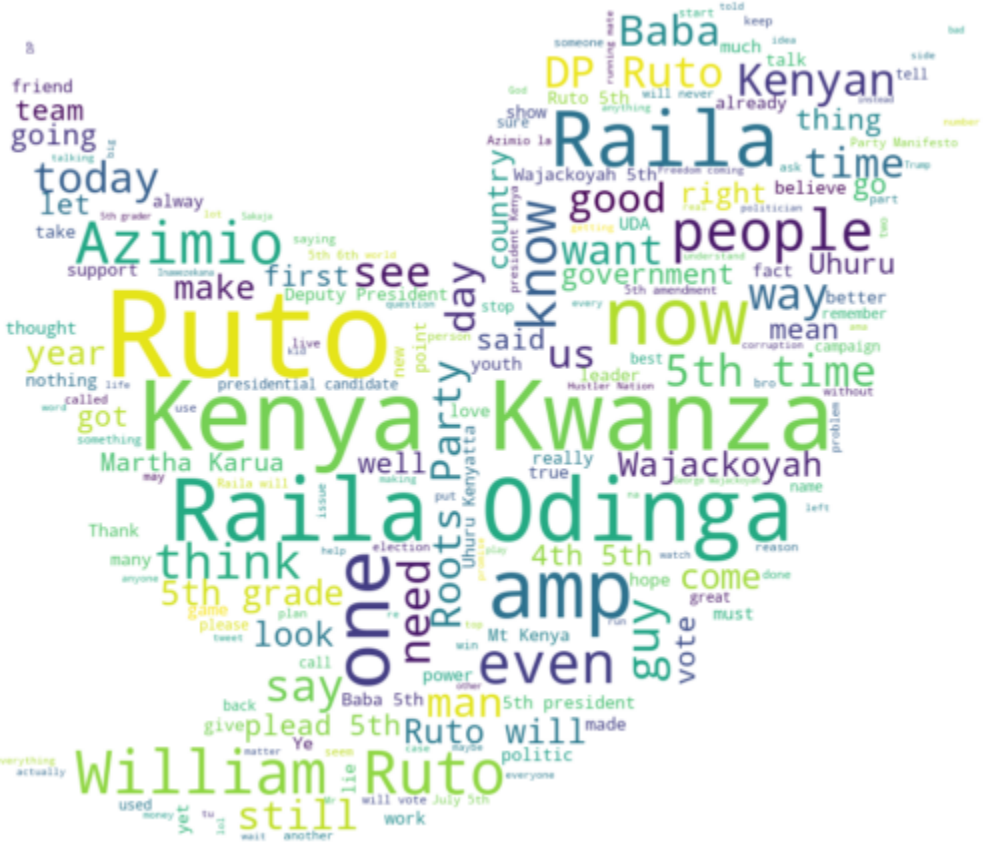
Most words used in negative tweets



Most words used in neutral tweets

Word Cloud for popularity

Further, a wordcloud was plotted to see which names or phrases appeared most in the tweets. This can be used to inform the tertiary companies i.e Ipsos on candidates' popularity.

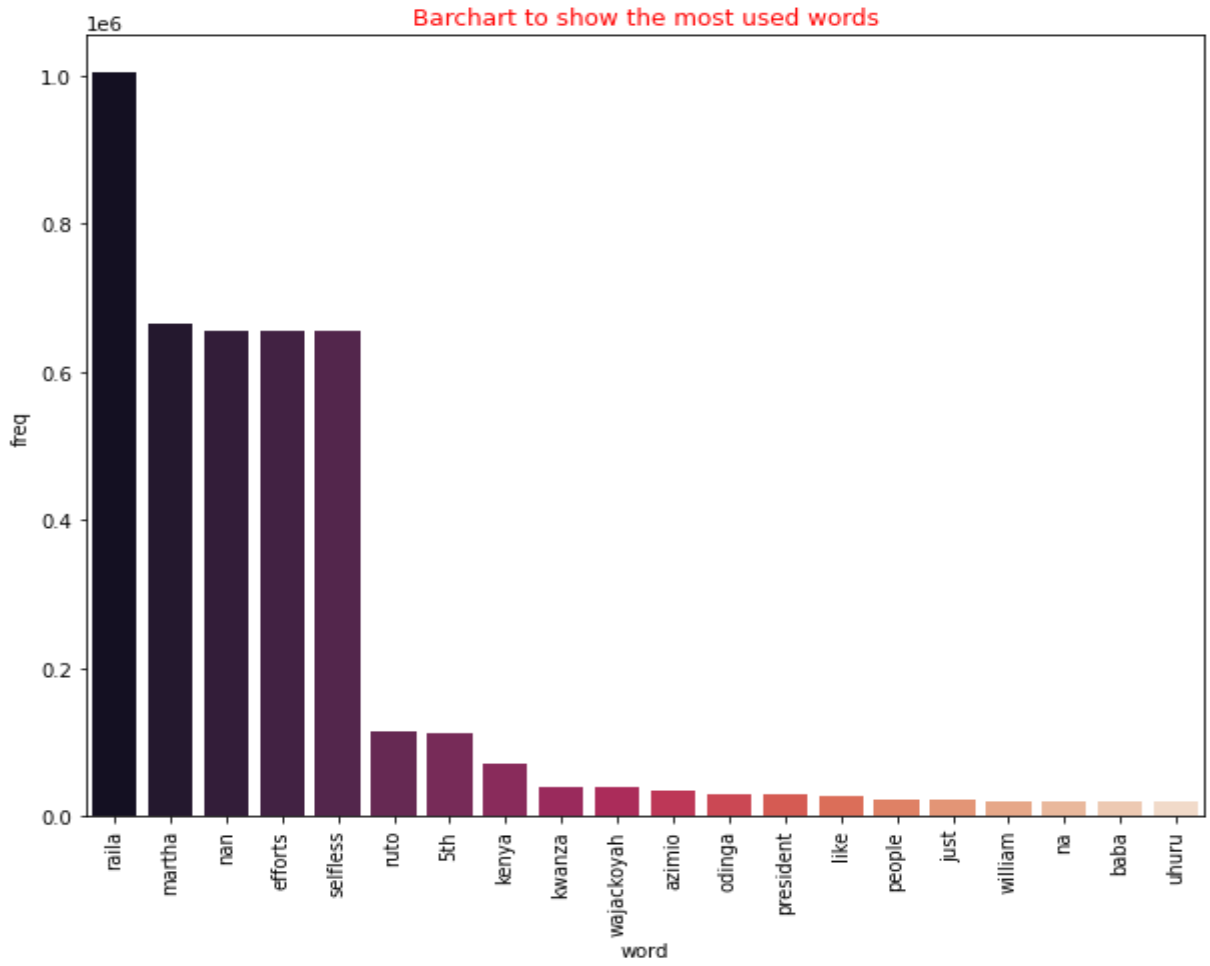


Among the ten most trending names or phrases of Kenyan political interest are the following:

- "Kenya Kwanza"
- "Raila Odinga"
- "Hustler nation"
- "William Ruto"
- "Martha Karua"
- "DP Ruto"
- "Baba 5th"
- "Wajackoyah 5th"
- "Moses Kuria"
- "Hon Raila"

Graph for the most used words in descending order.

A graph was also plotted to show the most used words. Below is the graph.

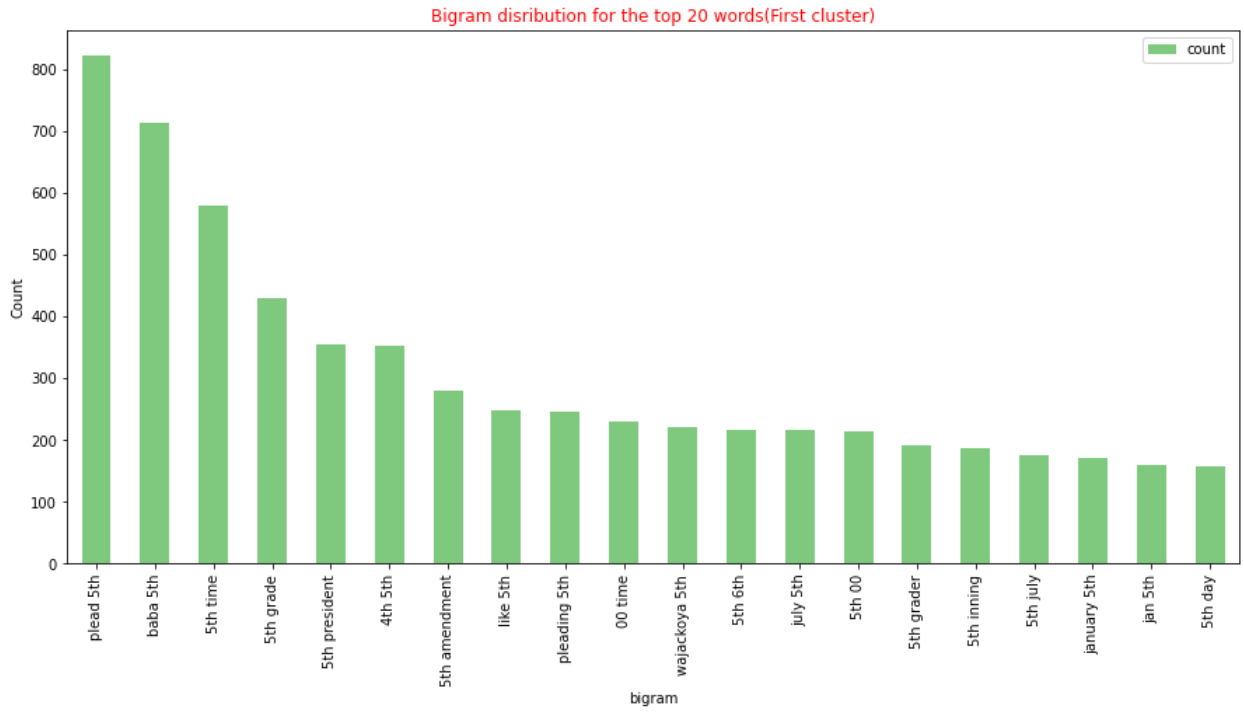


In the context of this project, the five most used words are “raila”, “martha”, “efforts” , “selfless” and “ruto”

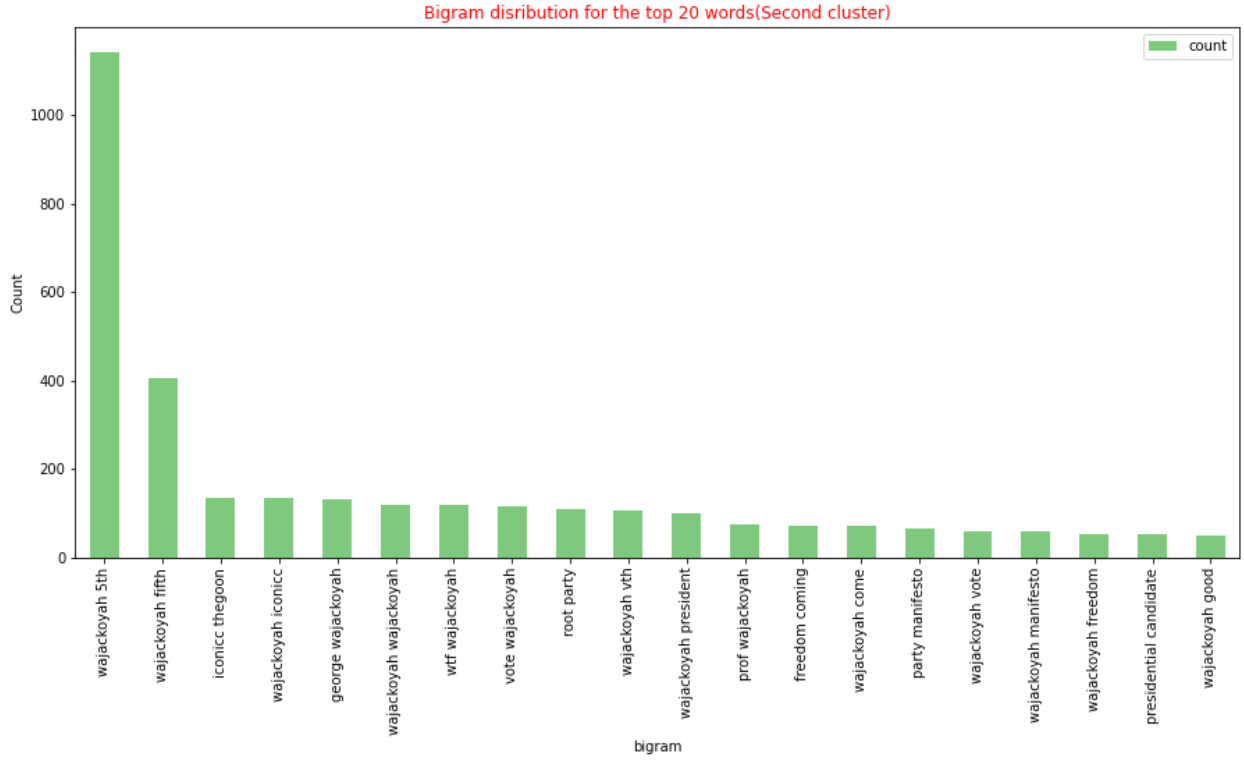
5.0 MODELING

In this section, K- means clustering was employed in which five clusters were used.

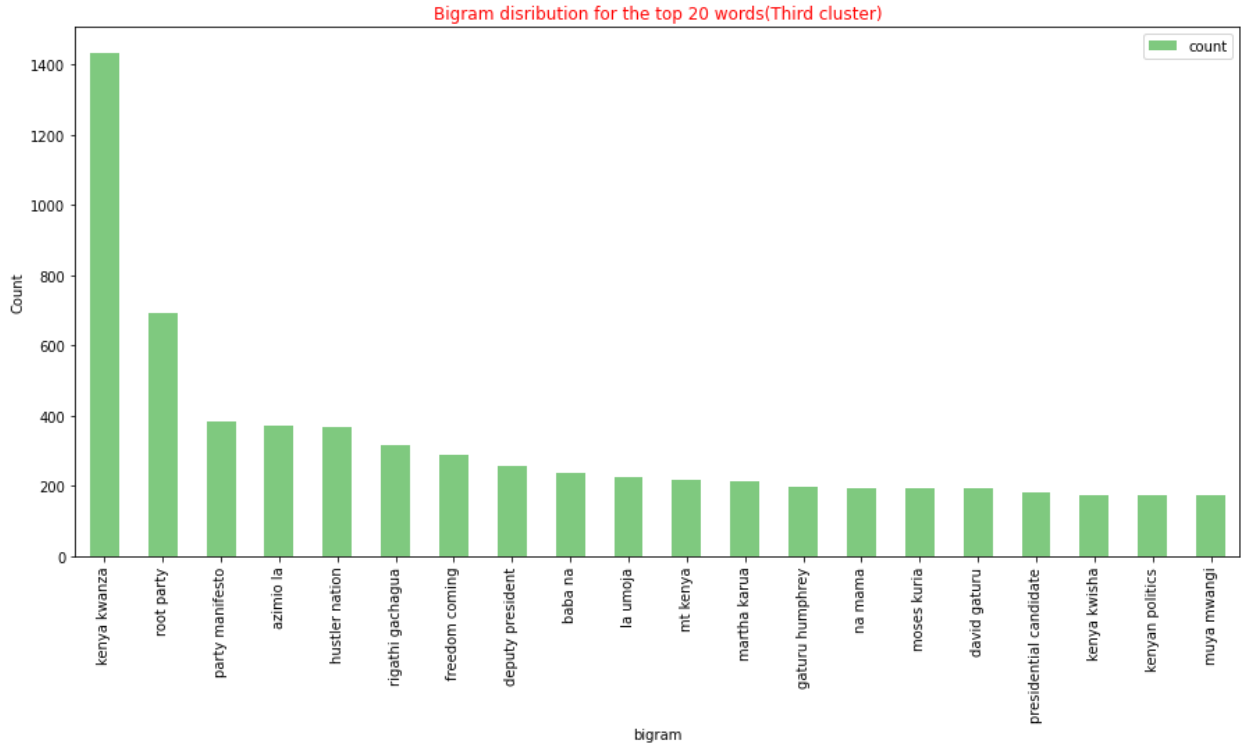
The first cluster captures the numerous mentions of '5th' which can be attributed mostly to irrelevant data and also the fact that we shall be voting in the fifth president of Kenya.



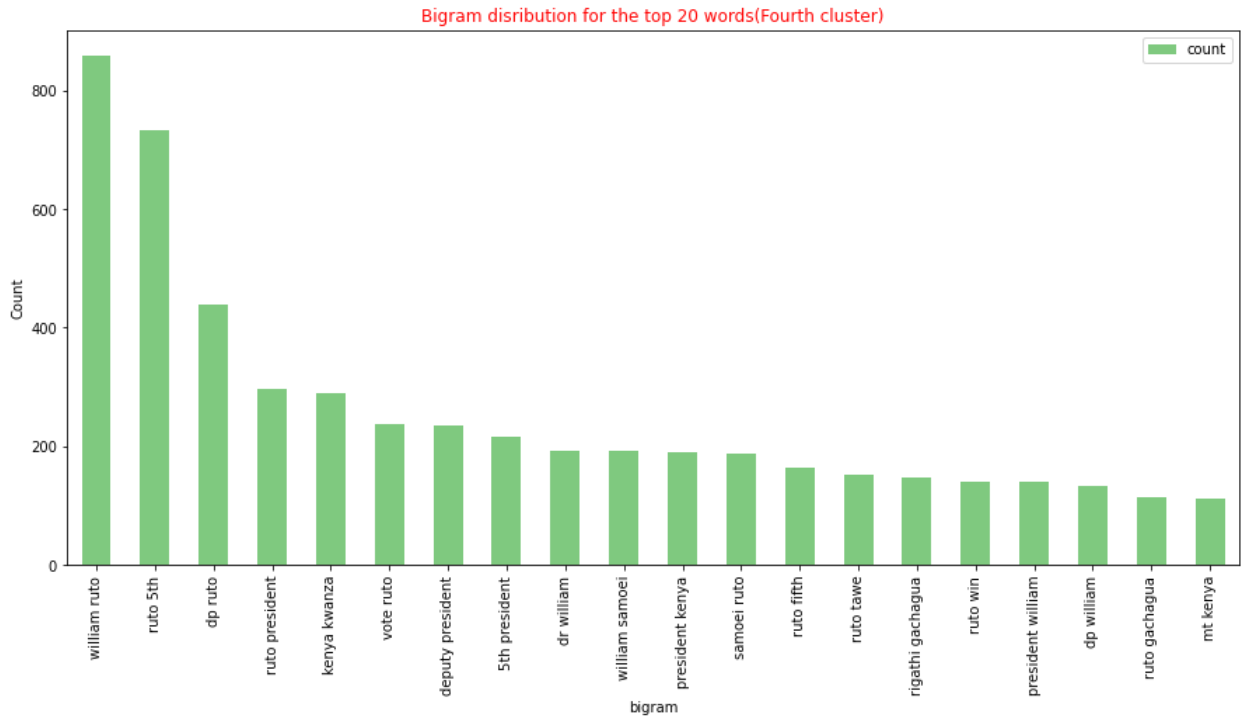
The second cluster contains captures events and mentions of presidential aspirant, George Wajackoyah.



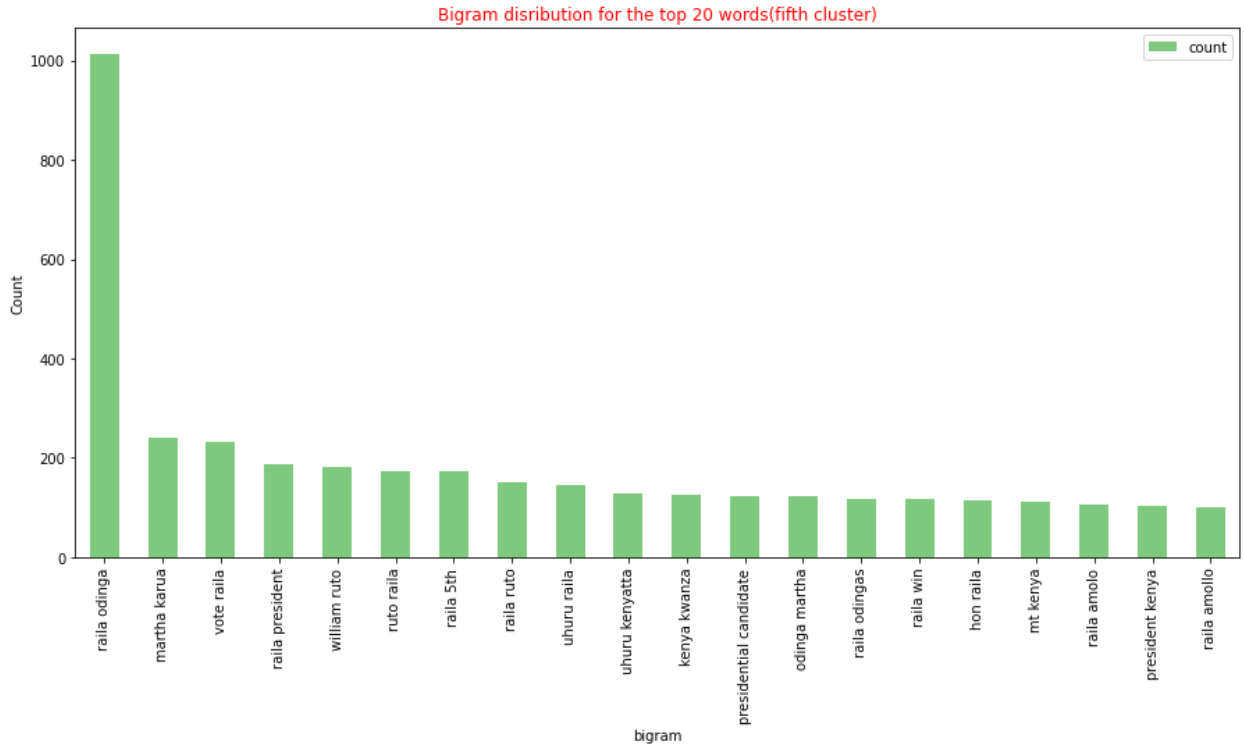
The third cluster includes events surrounding all other events other than the rest of the clusters.



The fourth cluster captures events surrounding presidential aspirant William Ruto.



The fifth cluster captures events surrounding Presidential aspirant Raila Odinga.



6.0 DEPLOYMENT

A platform was created using Streamlit, where a user can log in to their account and access information from the analysis. A person can access information about the most used words in Kenyans tweets and most likely winner for presidential polls.

7.0. CONCLUSION AND RECOMMENDATION

Conclusions

- From our analysis, we concluded that William Ruto is the most polarizing candidate. Ruto has the highest number of positive sentiments at the same time the highest number of negative sentiments
- We also concluded that Ruto is the most mentioned political figure in the country while Wajackhoyah is the least mentioned
- Factoring in neutral voters as undecided voters, it was concluded that a large part of the electorate is still undecided on who to vote for in the forth coming election.

- Without the undecided voters, from the analysis, it was concluded that William Ruto will most likely win the election in the first round with slightly over 50% of the votes.

Disclaimer

This prediction is made using only sentiments sourced from one social media platform hence is not fully representative of Kenya's electorate. Also, the data scrapped was for only 3 months and so cannot be used to represent all of Kenyan sentiments.
Also, Kenyan politics are not exactly issue-based but rather people get sentimental around certain leaders.

Recommendation

- Presidential candidates should continue to make themselves and their manifestos known to the public since there are still many undecided voters as per this analysis.