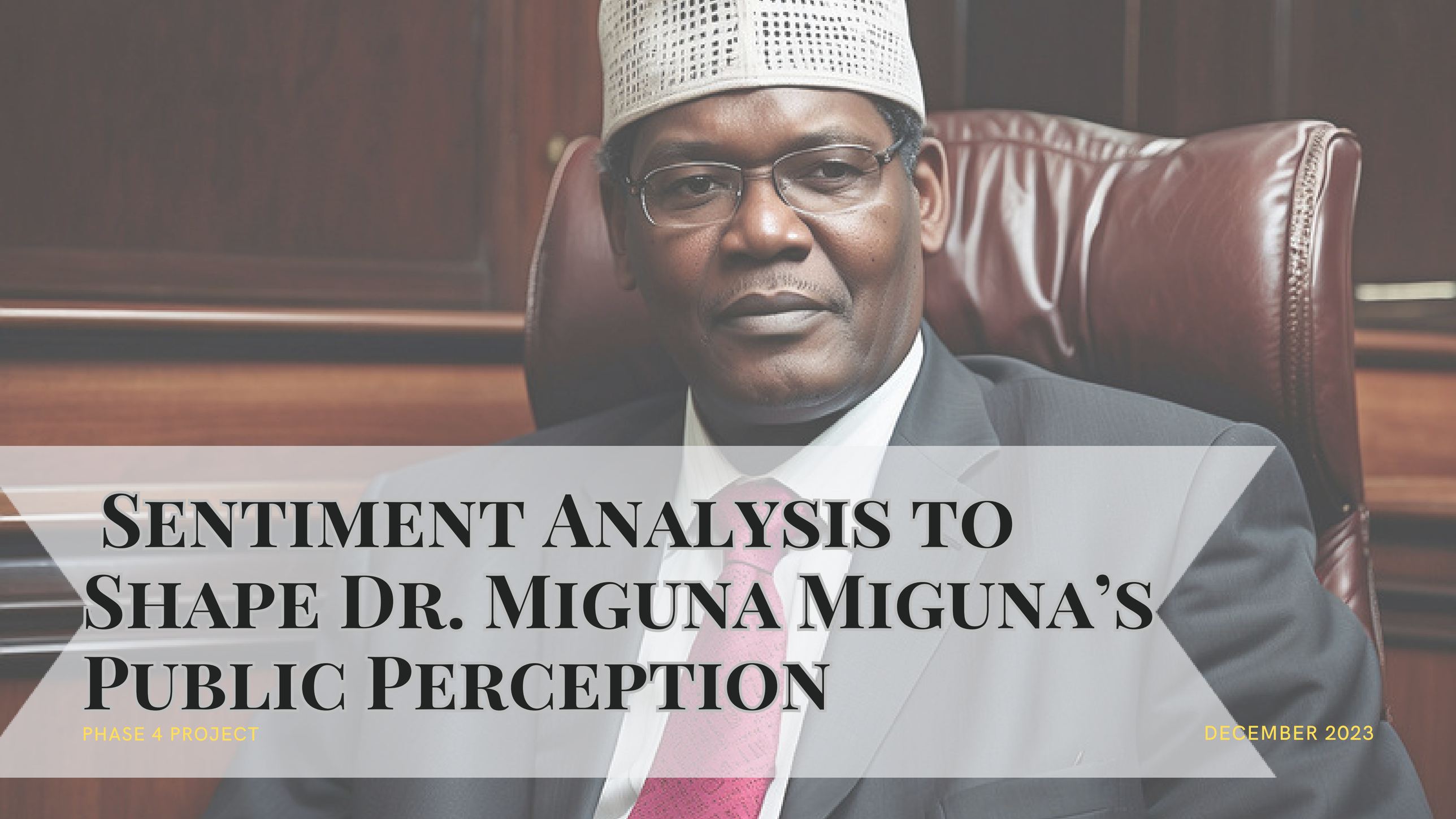


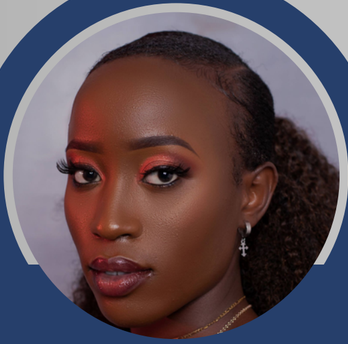
Sentiment Analysis to Shape Dr. Miguna Miguna's Public Perception

PHASE 4 PROJECT

DECEMBER 2023

PREDICTIVE PRODIGIES:

Richard Taracha



Juliet Wanja

Eva Kiio

PROJECT OVERVIEW

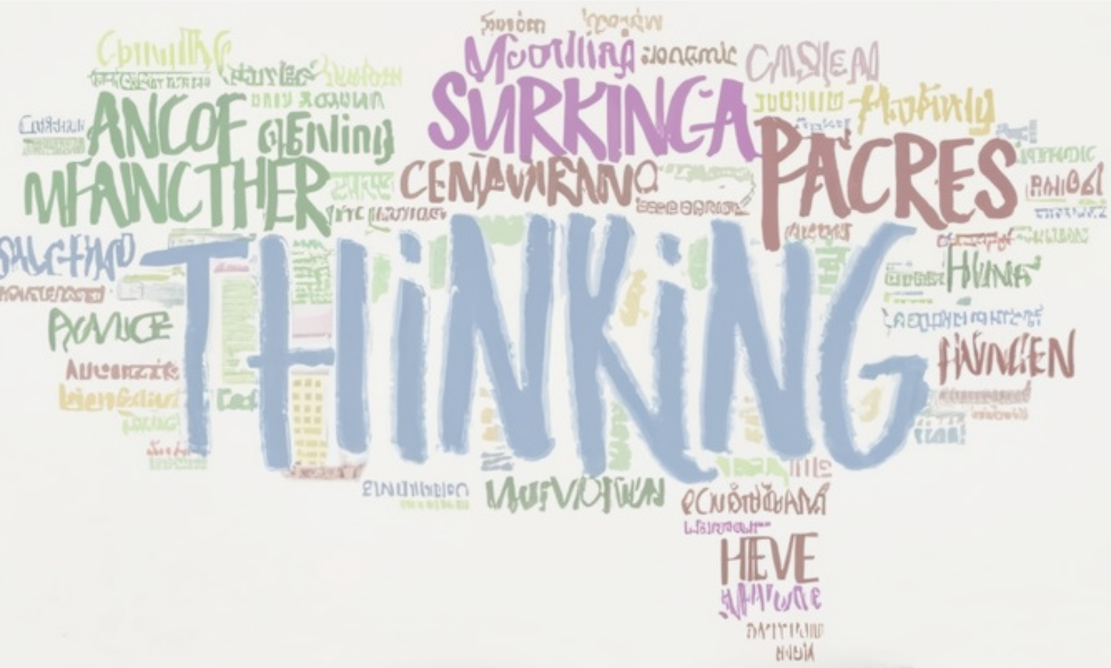
PREDICTIVE PRODIGIES

INTRODUCTION

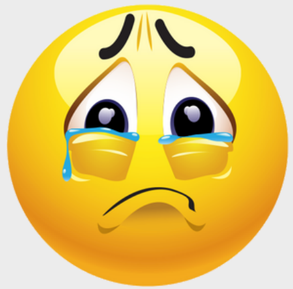
- In the dynamic world of politics, public perception plays a crucial role in shaping the success or failure of a political figure.

- Effectively managing and understanding the vast amount of data generated on Twitter can be a daunting task. This is where Twitter sentiment analysis using NLP (Natural Language Processing) emerges as a powerful solution.

what is sentiment analysis?



Sentiment analysis is a technique used to identify, extract, and quantify emotions, opinions, or attitudes expressed in text data.



The importance of Sentiment Analysis in Understanding Public Opinion

- Sentiment analysis plays an essential role in helping us understand the collective opinion, feelings, and views each individual expresses on social media platforms.

- It helps in determining how the public feels, brings out trends, and realizes the effect of social media personalities upon their followers.

- **Natural Language Processing** it is a branch of artificial intelligence
- It enables computers to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant.

- **Project focus:** Analyzing Miguna Miguna's 2019-2022 tweets using NLP and ML to discern public sentiment and message resonance.

OBJECTIVE

Key Activities

1. **Data Preprocessing**
2. **EDA**
3. **Modeling**
4. **Evaluation**
5. **Deployment**

DEVELOP A MACHINE LEARNING MODEL TO ACCURATELY DISCERN PUBLIC SENTIMENT FROM MIGUNA MIGUNA'S TWEETS

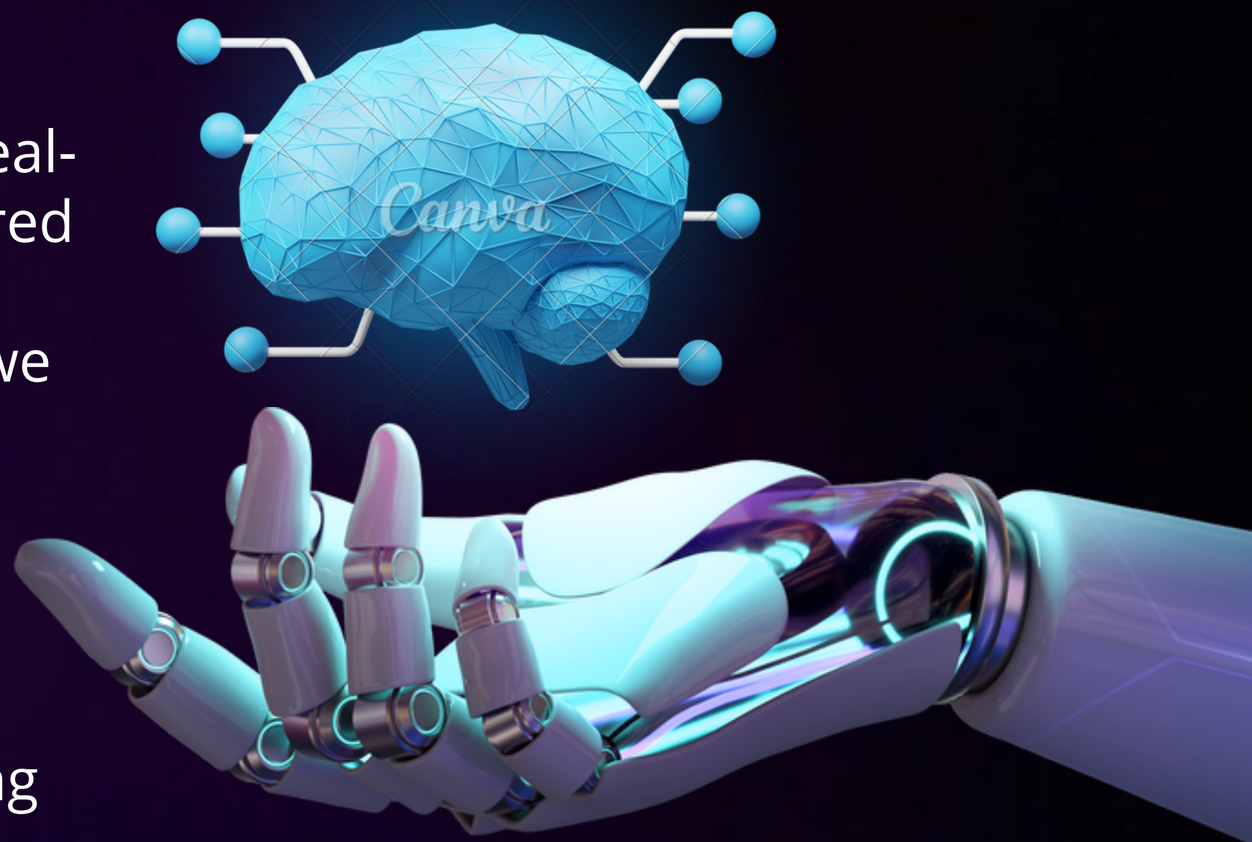
Results

Offering insights into audience perception and message resonance over time.

DATA COLLECTION

- Twitter's API provides incredible access to real-time data, but for this project, we encountered limitations due to API access restrictions for historical tweets. To overcome this hurdle, we turned to a powerful third-party tool called Twint.

- One of the key advantages of Twint was its ability to retrieve tweets without direct API access, allowing us to collect tweets spanning several months.



DATA UNDERSTANDING

Data collected contains 43,478 entries.

The four columns of interest include:

- **ID**: Each entry has a unique identifier.
- **Date**: The date of the tweet's publication.
- **Username**: The username of the individual that posted the tweet.
- **Tweet**: The text content of the tweet itself.

DATA PREPROCESSING

1. Text Cleaning

- The first step was to preprocess textual data so as focus solely on the text that conveys sentiment

- The **NeatText Library** was utelized for the removal of :

Mentions / User handles
@migunamiguna

Hashtags (#)

URLs (https://)

Special Characters
(!?>.:)

Whitespaces
(" ","\t"...)

Emojis

Contractions
(I'm: contraction for "I am")

Stopwords
(a,an,the,in,..)

2. Linguistic Processing

Tokenization:

Dividing text into smaller units (tokens) like words or phrases.

01

02

Parts of Speech Tagging:

Identifying grammatical elements in text (nouns, verbs, etc.).

Lemmatization:

Reducing words to their base or root form for analysis.

03

04

Sentiment Calculation:

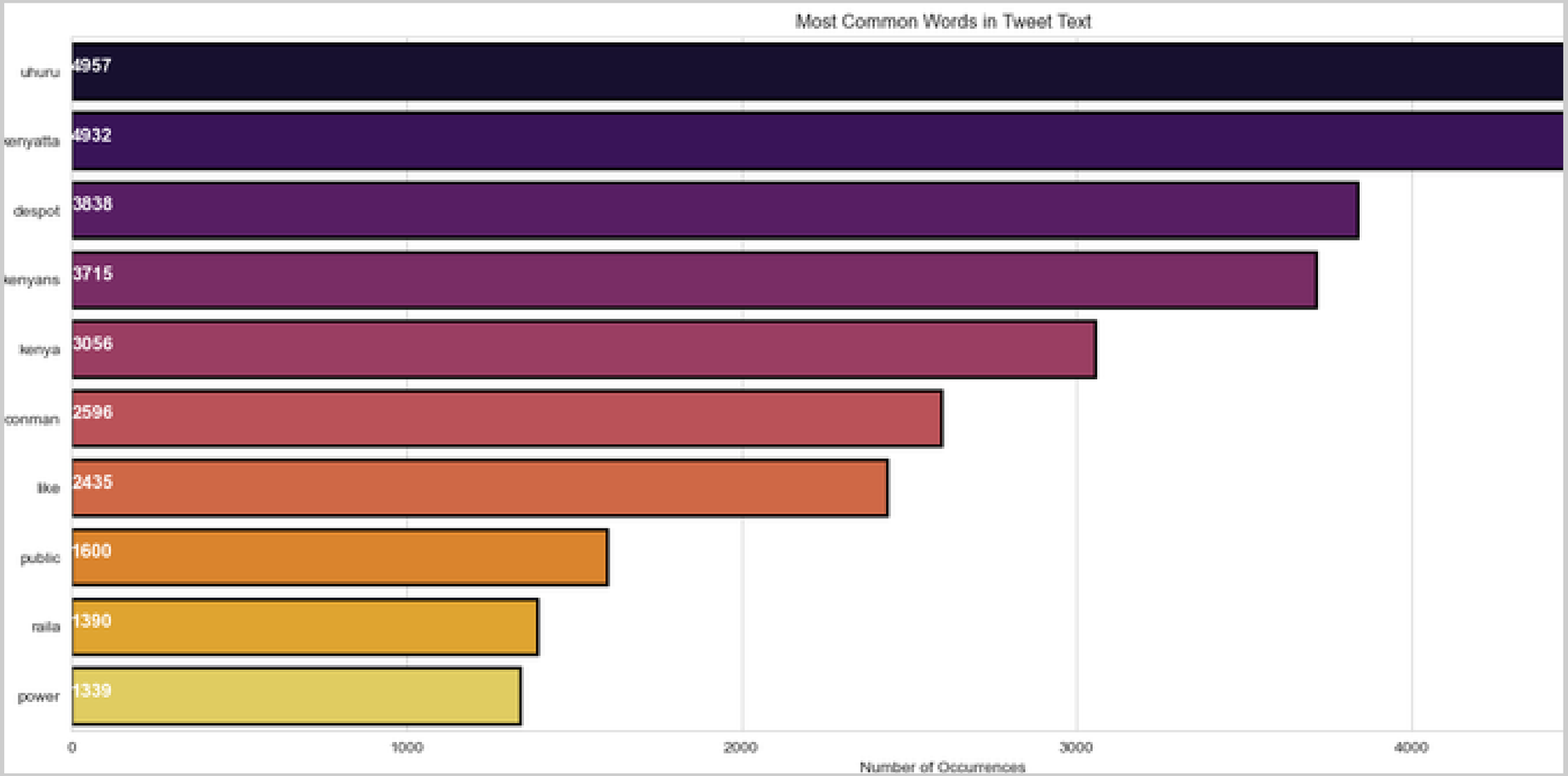
Assessing polarity (positive/neutral/negative) and subjectivity of text for sentiment analysis.

EDA

EXPLORATORY DATA ANALYSIS

EDA

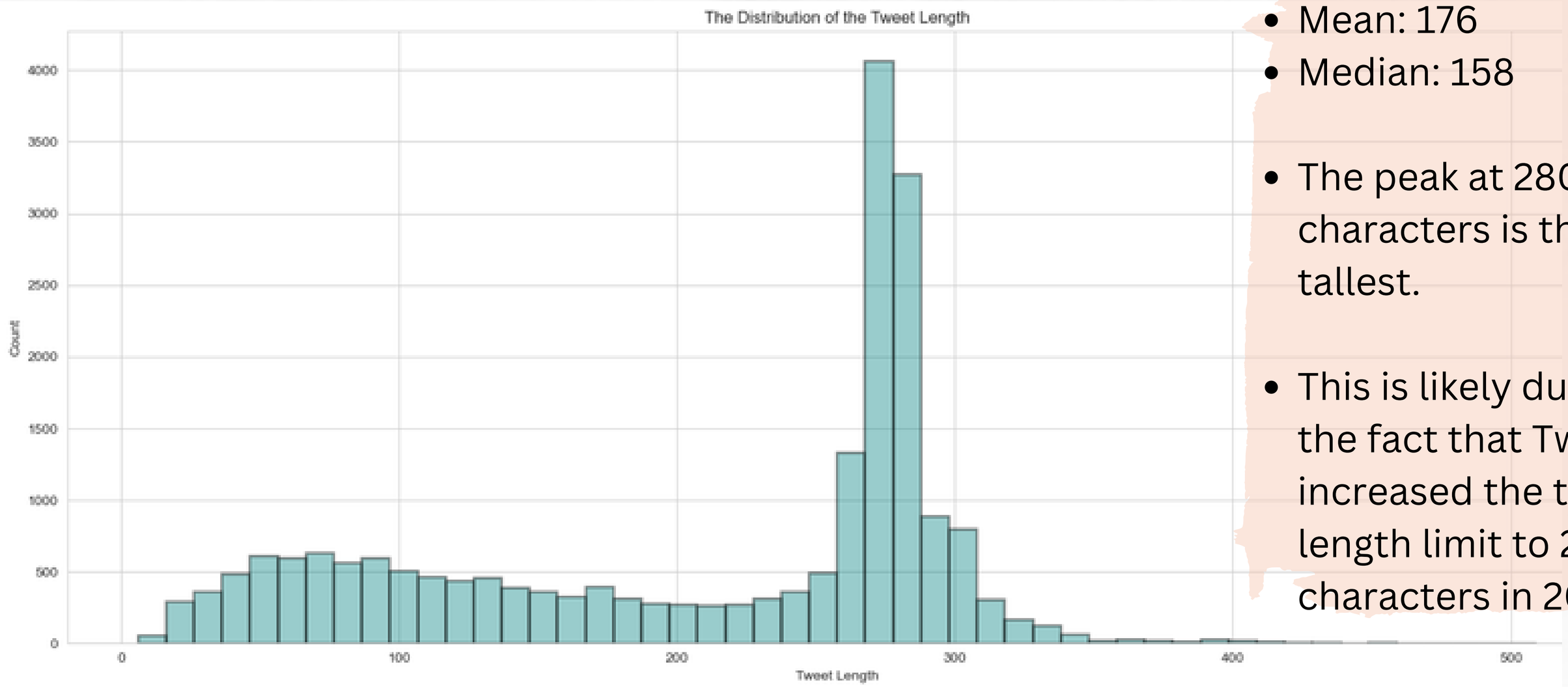
Word Count



The chart presents the most frequent terms from Miguna Miguna's tweets, revealing recurring themes such as 'Uhuru,' 'Kenyatta,' 'despot,' and 'Kenyans,' offering insight into prevalent topics in the discourse."

Character Count

- The graph shows a right-skewed distribution of tweet lengths, with a long tail of tweets over 300 characters.

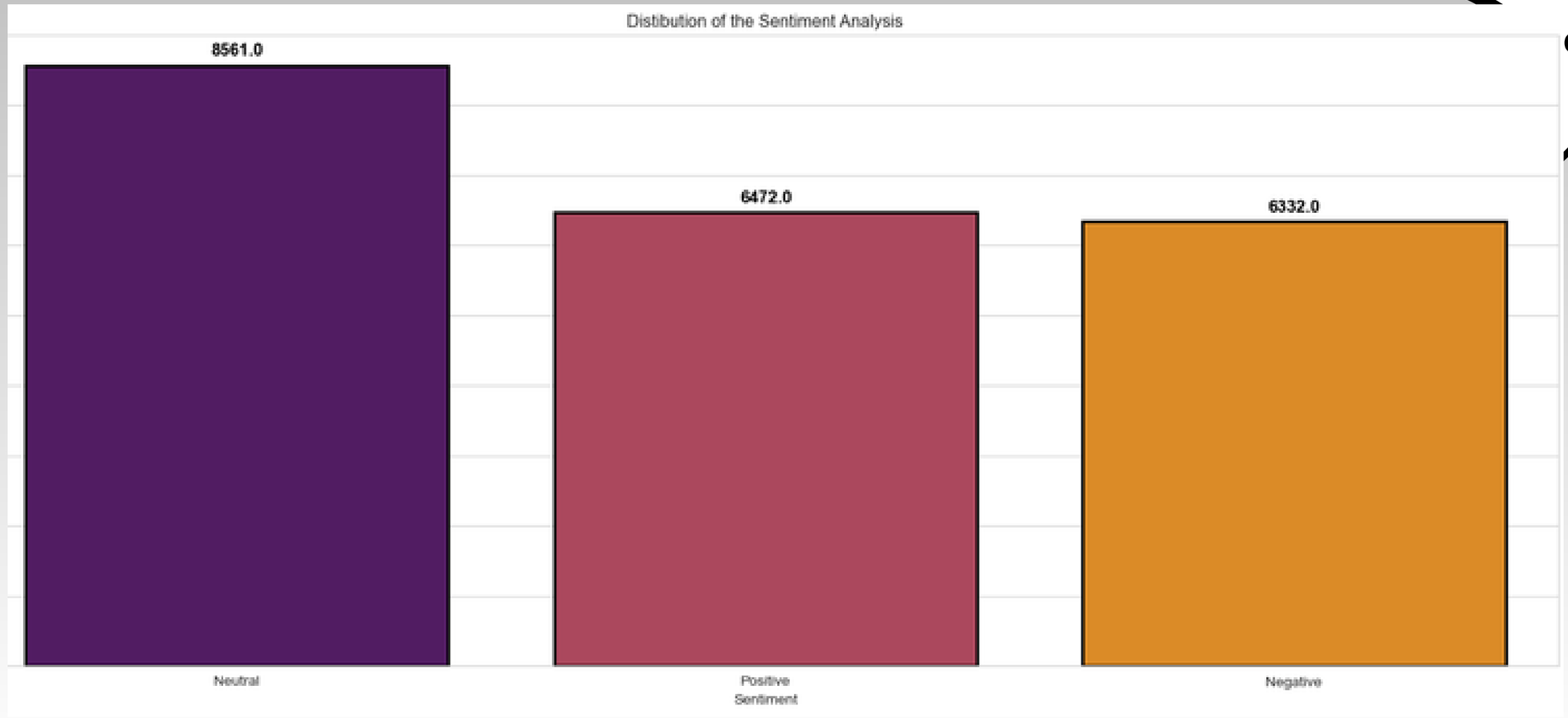


- Mean: 176
- Median: 158

- The peak at 280 characters is the tallest.

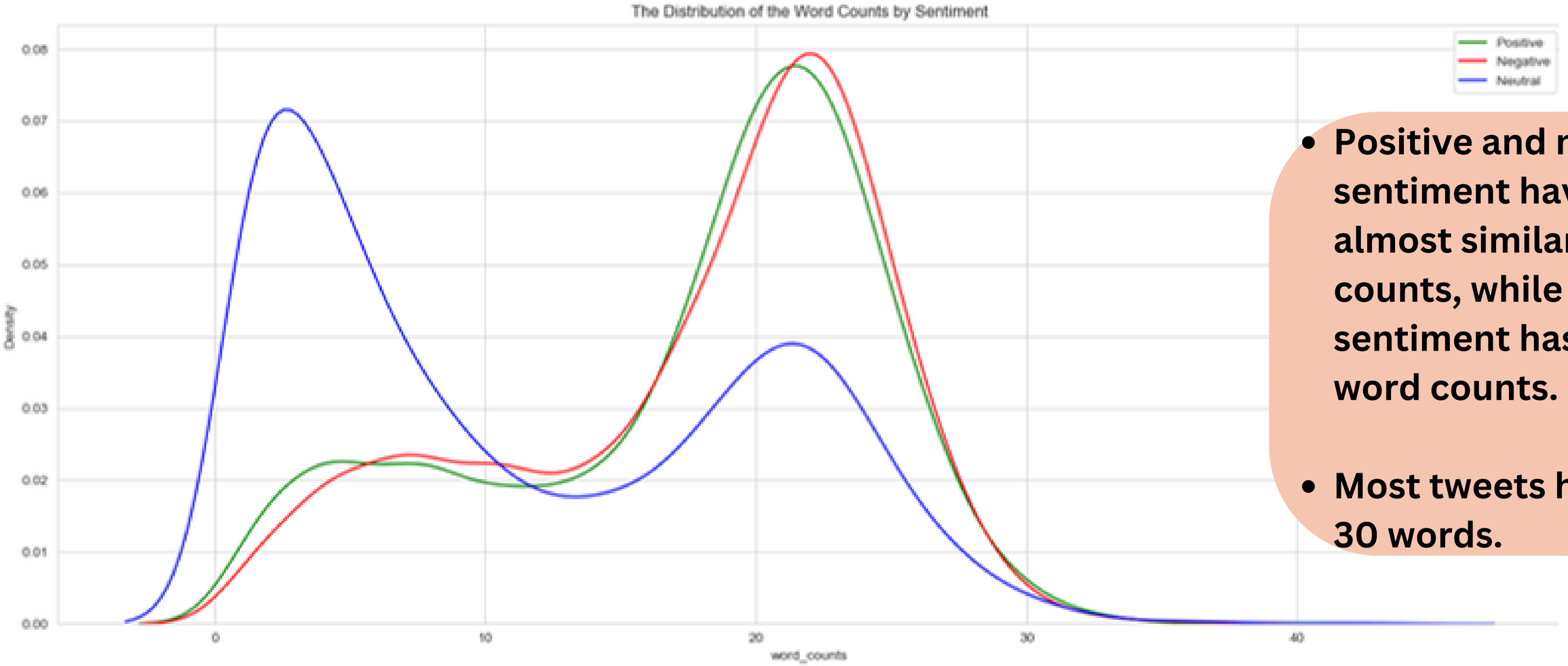
- This is likely due to the fact that Twitter increased the tweet length limit to 280 characters in 2017.

Distribution of Sentiments



The graph shows that the most frequent sentiment among the data is neutral, followed by positive and negative. This suggests that the data is balanced and diverse in terms of opinions and emotions.

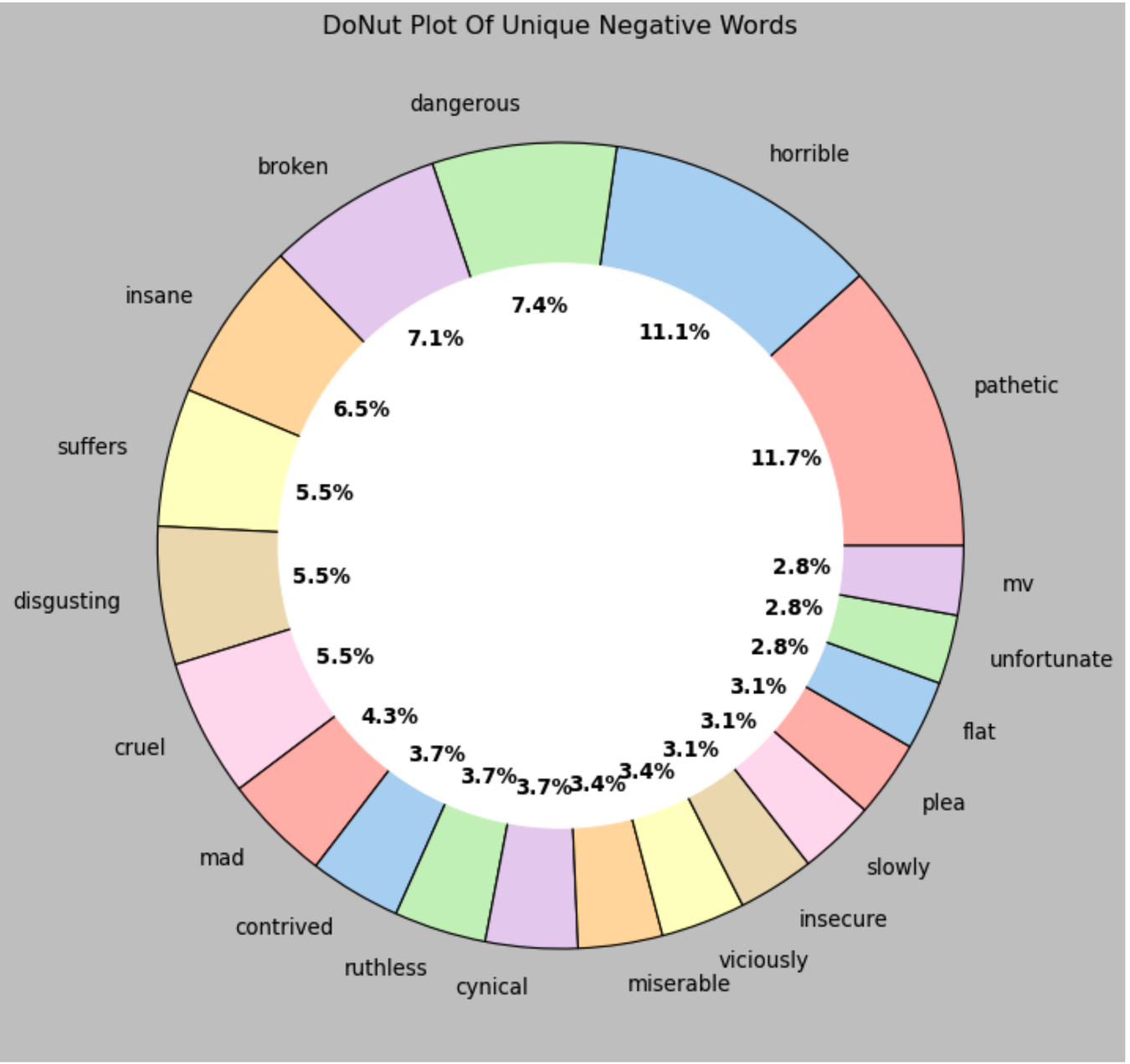
Word Count by Sentiment



- **Positive and negative sentiment have almost similar word counts, while neutral sentiment has lower word counts.**
- **Most tweets have 20-30 words.**

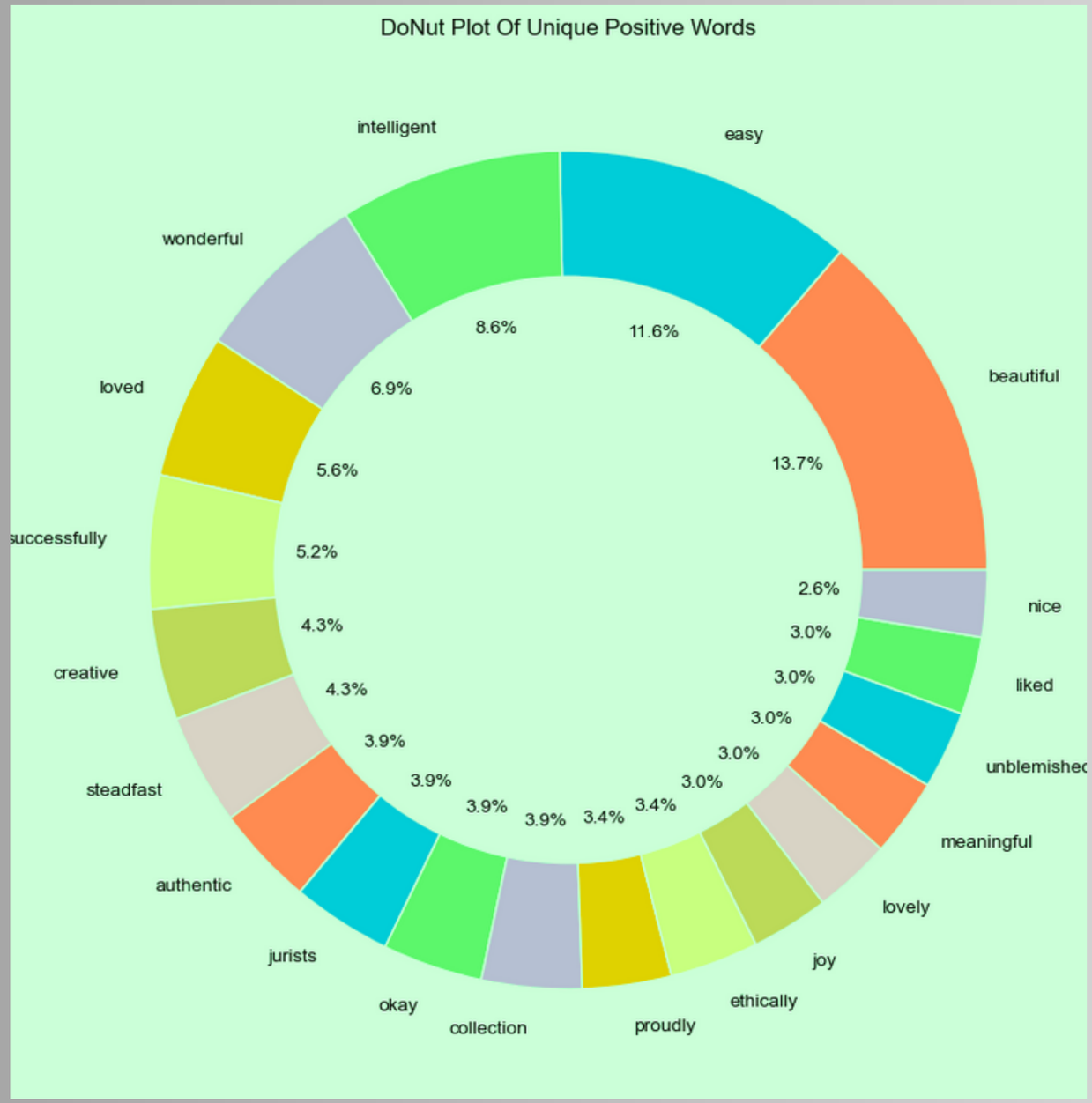
The length of the text does not affect the polarity of the sentiment much. This suggests that people are just as likely to express positive or negative emotions in a short tweet as in a long one.

Most Unique Negative Words



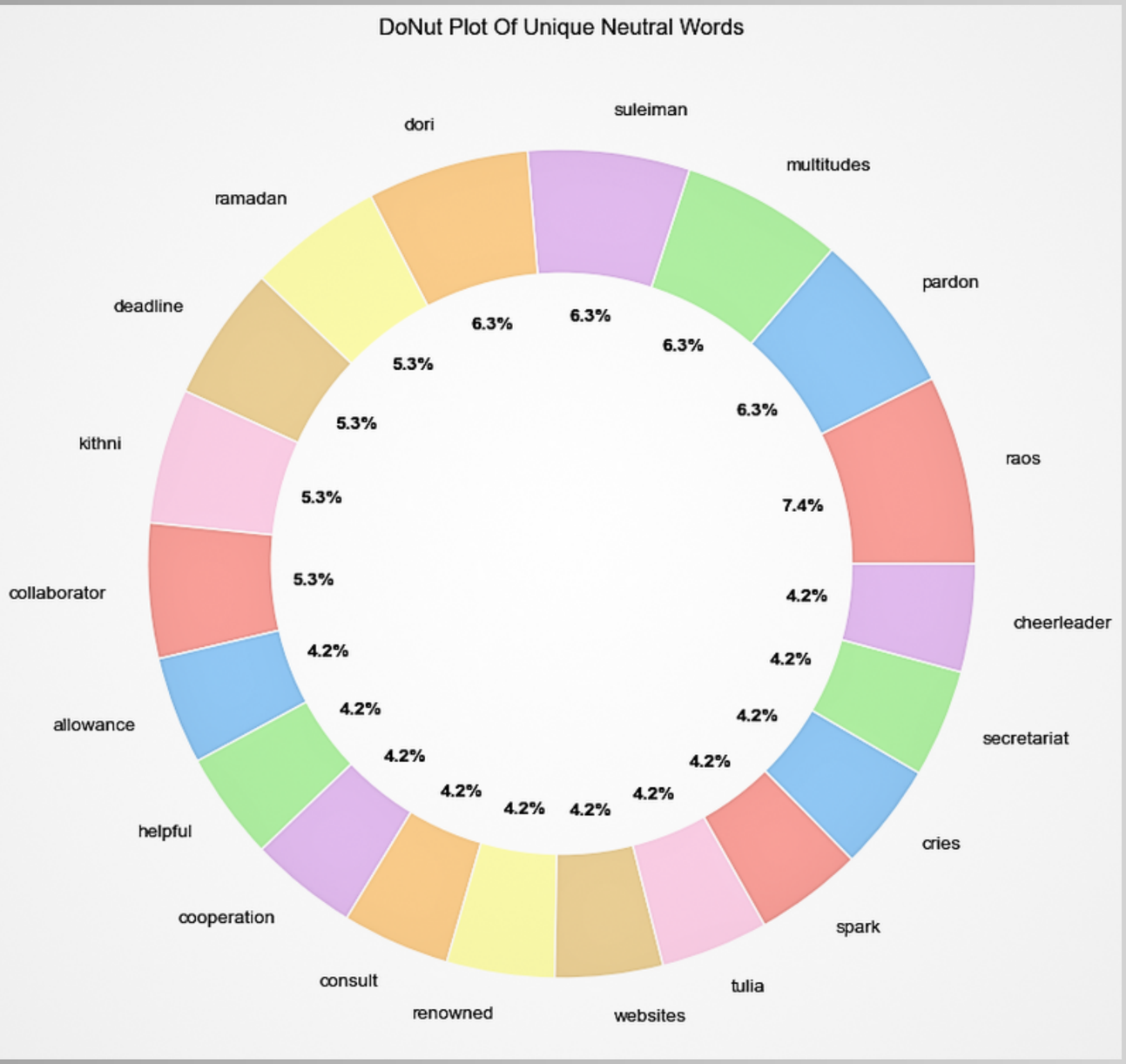
The donut chart shows the percentage of unique negative words in Miguna Miguna's twitter space

Most Unique Postive Words



The donut chart shows the percentage of unique positive words in Miguna Miguna's twitter space

Most Unique Neutral Words



The donut chart shows the percentage of unique neutral words in Miguna Miguna's twitter space

MODELING

Modeling Approach:

Logistic Regression

- it is aLinear model that works well for binary classification problems like sentiment analysis.
- it Estimates the probability that a given input belongs to a particular class (positive/negative sentiment).
- it is an Interpretable model, making it easier to understand the impact of different features on predictions.
- Can handle large feature spaces efficiently.

Complement Naive Bayes (CNB)

- Variation of Naive Bayes designed to handle imbalanced datasets (when classes are unevenly distributed).
- Particularly effective for text classification tasks like sentiment analysis.
- Accounts for imbalanced class distributions by adjusting the probability estimation.
- Less impacted by the presence of frequent words in one class but not the others.
- Works well with text data containing a large number of features (words).

Random Forest

- Ensemble learning method that builds multiple decision trees during training.
- Each tree votes for the most popular class, and the final prediction is the majority vote.
- Handles non-linear relationships well and can capture complex interactions in the data.
- Less prone to overfitting compared to individual decision trees.
- Capable of handling large datasets with high dimensionality.

Results and Findings:

logistic regression

Training Score: 1.0 Test Score:0.91

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.92	0.89	0.91	1324
1	0.90	0.95	0.92	1746
2	0.90	0.87	0.89	1203
accuracy			0.91	4273
macro avg	0.91	0.90	0.91	4273
weighted avg	0.91	0.91	0.91	4273

Complement Naive Bayes (CNB)

Training Score: 0.87 Test Score:0.69

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.64	0.81	0.72	1324
1	0.83	0.54	0.65	1746
2	0.63	0.78	0.70	1203
accuracy			0.69	4273
macro avg	0.70	0.71	0.69	4273
weighted avg	0.72	0.69	0.69	4273

Tuned Random forest

Training Score: 0.86 Test Score:0.78

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.89	0.43	0.58	1324
1	0.56	0.99	0.71	1746
2	0.88	0.40	0.55	1203
accuracy			0.65	4273
macro avg	0.78	0.60	0.61	4273
weighted avg	0.75	0.65	0.62	4273

Logistic Regression

Training Score: 1.0 Test Score:0.91

CLASSIFICATION REPORT

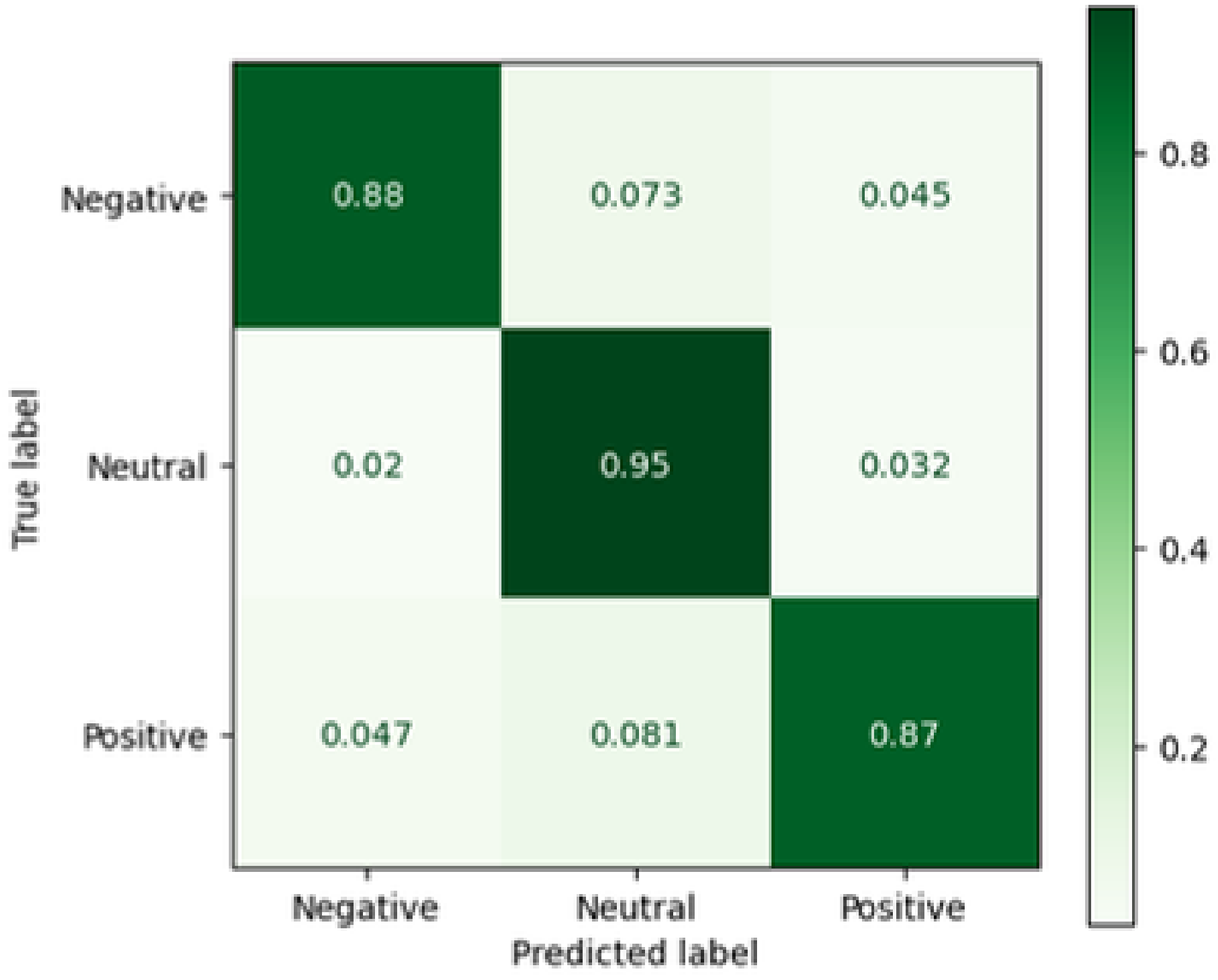
---

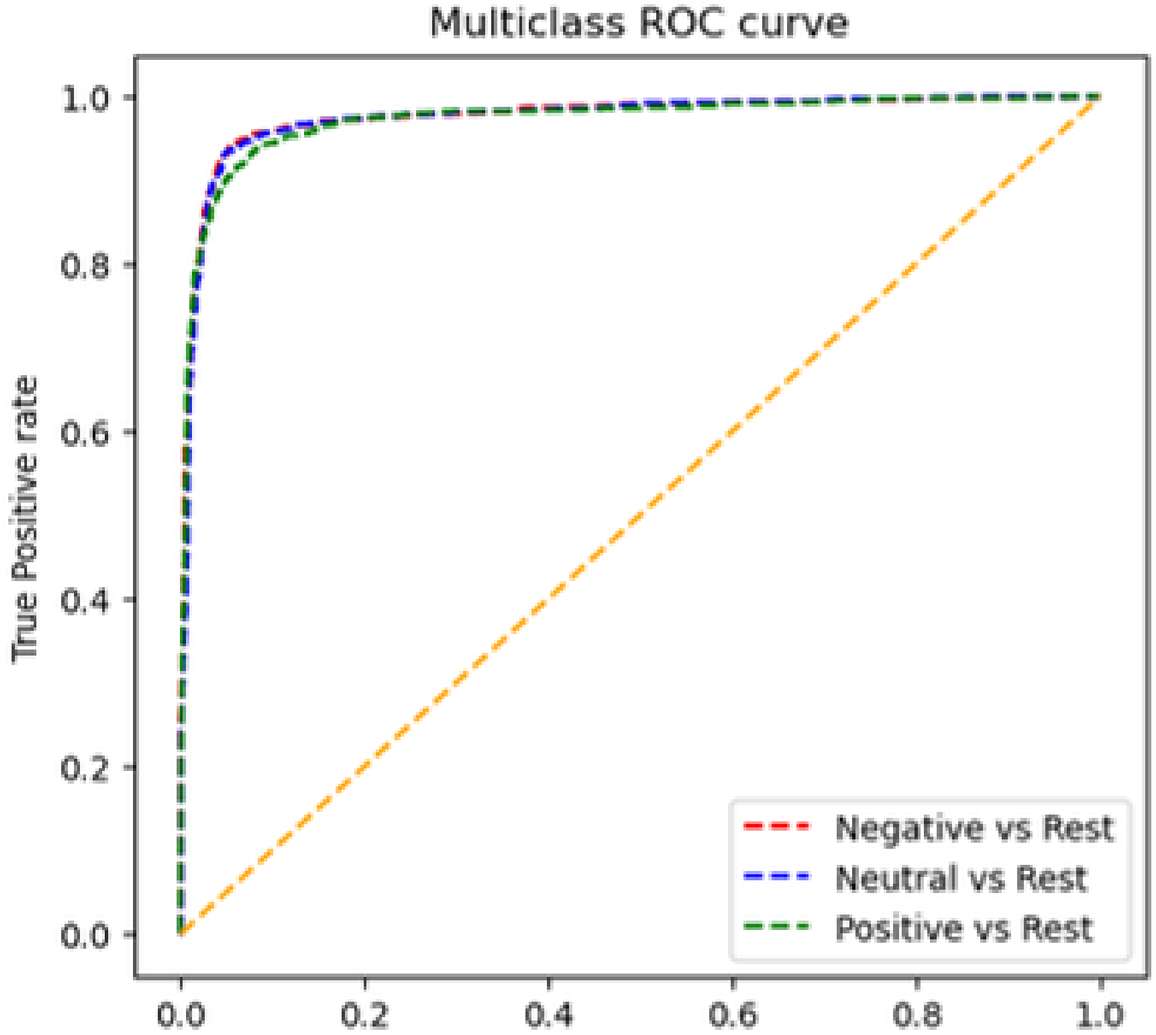
	precision	recall	f1-score	support
0	0.92	0.89	0.91	1324
1	0.90	0.95	0.92	1746
2	0.90	0.87	0.89	1203
accuracy			0.91	4273
macro avg	0.91	0.90	0.91	4273
weighted avg	0.91	0.91	0.91	4273

The fine-tuned logistic regression model emerged as the top performer in this task, showcasing an impressive macro recall score of 90%. we can now implement our model in the deployment phase.

Tuned Logistic Regression Model

Below are the results of the best-performing model, showcasing its strong classification and discriminatory abilities through the confusion matrix and ROC curve.





Top model:

Impressive 0.90 recall-macro-avg score.

F1-scores range

89-92%, indicating accurate predictions.

"Neutral" class

93%, outperforming other models.

"Negative" and "Positive" classes

10% increase to 88% from 78%.

Test accuracy

Highest recorded at 91%, surpassing previous models.

CONCLUSION

To conclude, Dr. Miguna Miguna emerges as an outspoken and opinionated figure in Kenyan politics. He engages deeply with discussions on leadership and societal issues, shaping his presence on Twitter through critical viewpoints, a clear communication style--balanced yet strongly expressed sentiments that reflect a passionate stance for matters close to him.

RECOMMENDATION

1 Strategies for Enhancing Engagement:

Through posing questions, conducting polls, and actively seeking feedback from our audience: we aim to comprehend their perspectives and needs--a strategy that fosters trust; it establishes an essential connection with them.

2 Approach to Constructive Criticism:

Please offer counsel: in your criticisms of Kenyatta and Uhuru, maintain a tone that's both respectful and constructive; eschew inflammatory or abusive language—not only to sidestep potential legal entanglements but also to uphold credibility.

3 Positive Aspects Acknowledgement:

While you suggest acknowledging the positive aspects or achievements of Kenyatta and Uhuru, remember to offer suggestions for their improvement: this approach showcases fairness and a willingness to collaborate—both essential elements in striving towards Kenya's betterment.

phase 4
project

Thank You

Connect with us.

@the predictive prodigies

@thepresictiveprodigies.com