PROTESTS

A DATA SCIENCE APPROACH

Group 9 Capstone Project, DSFPT06

Group Members:
Magdalene Ondimu
Brian Kariithi
Najma Abdi
Leon Maina
Wilfred Lekishorumongi

PREDICTING STATE RESPONSES TO PROTESTS

A Machine Learning Approach Using Topic Modelling, Sentiment Analysis, and Classification Using Key Protest Attributes

OVERVIEW

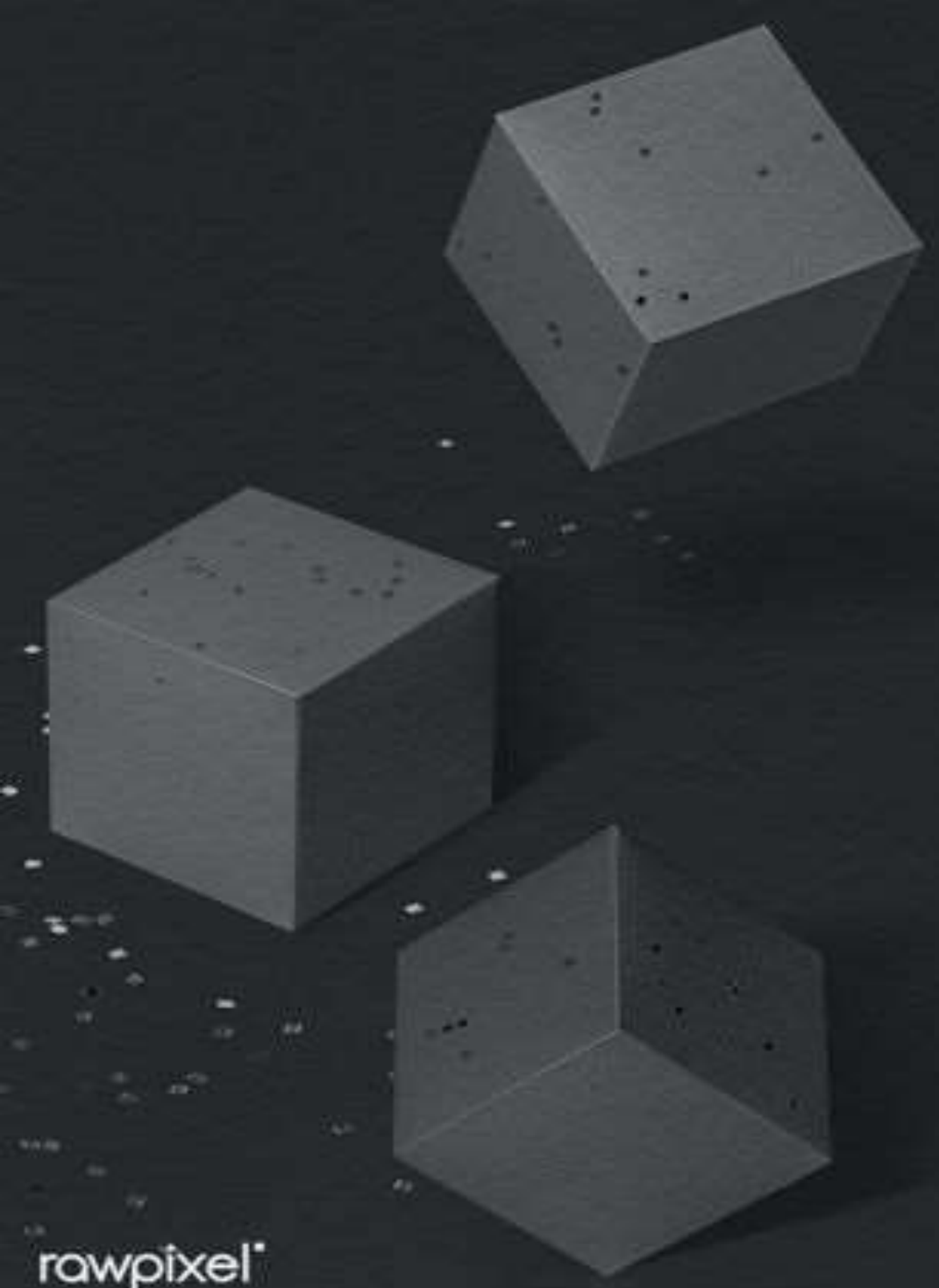
The project aims to analyze global protest events to understand their underlying causes, geographical and temporal trends, patterns, and state responses.

This comprehensive analysis is crucial for informing policy and decision-making to manage socio-political unrest effectively.

PROBLEM STATEMENT

Understanding and predicting state responses to protests is crucial for improving state-citizen relations and managing social unrest.

This project aims to provide valuable insights into public sentiment during protests and help policymakers and organizations make informed decisions. By leveraging machine learning, we can uncover patterns and trends in protest-related data, enabling more proactive and effective responses.



MAIN OBJECTIVES

- Develop and evaluate machine learning models to predict state responses to protests based on the identified sentiments, topics and other protest attributes such as region, protest duration and protest identity.
- Conduct sentiment analysis on protest notes and tweets to classify sentiments as positive, negative, or neutral.
- Identify and summarize key topics from protest-related communications.

SPECIFIC OBJECTIVES

- Enhance data preprocessing and feature engineering techniques to improve model performance.

- Perform dimensionality reduction to handle high-dimensional data and reduce overfitting.

- Fine-tune hyperparameters and explore ensemble methods to optimize model accuracy.

DATA COLLECTION

Description:

- Collected data from the Havard mass mobilization dataset https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/HTTWYL/TJJZNG
- Collected and Analyzed Tweets from Recent Protest Activities (July 13th to July 23rd 2024)

Data Cleaning: Handled missing values, removed duplicate values, filled null values, converted textual data to numerical data among others.

DATA UNDERSTANDING

Dataset Overview:

- **Source**: The dataset is sourced from the Mass Mobilization Project at Harvard, which records global instances of protests and state responses.

- **Time Period**: Covers a wide range of years from 1990-2020, providing a longitudinal view of protest dynamics.

- **Event Details**: Includes event dates,size of the protest,countries and regions of protests.

Tweets Key Features :

- **Date**: The date when the tweet was posted.

- **Sentiment:** The sentiment expressed in the tweet, which could be positive, negative, or neutral.

CURRENT PROTESTS IN KENYA

Protesters hold a Kenyan flag outside the Kenyan Parliament after storming the building during a nationwide strike to protest against tax hikes and the Finance Bill 2024 in downtown Nairobi, on June 25, 2024.

Protesters make signs with their arms in front of Kenya police officers during a demonstration against tax hikes as Members of the Parliament debate the Finance Bill 2024 in downtown Nairobi, on June 18, 2024.

CURRENT PROTESTS FROM AROUND THE WORLD

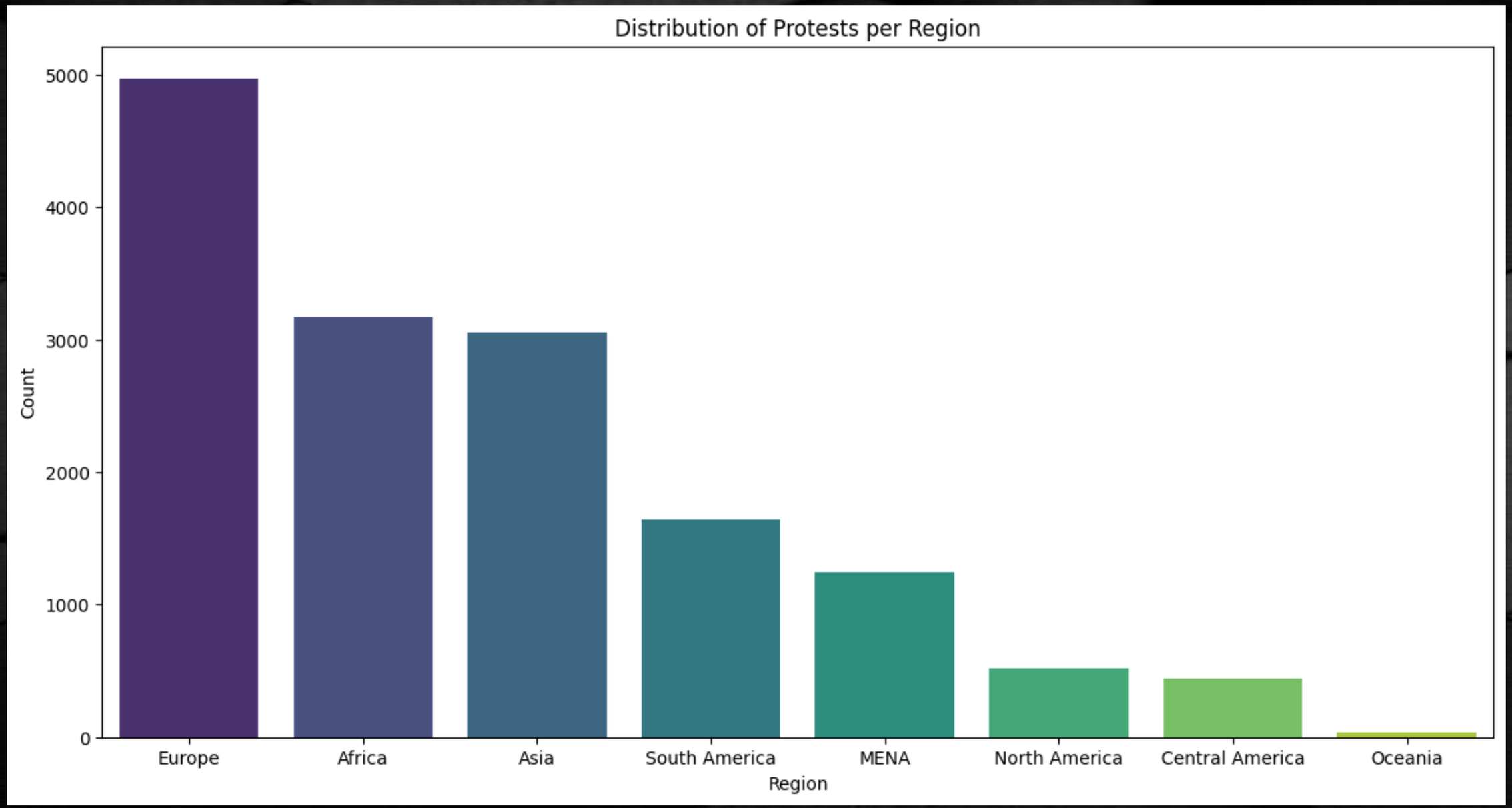
A protestor in Toronto, Canada carrying a placard highlighting a pursuit for freedom of the Palestinians and a stop to genocidal killings.

Anti-government protesters display Bangladesh's national flag as they storm Ex Prime Minister Hasina's palace.

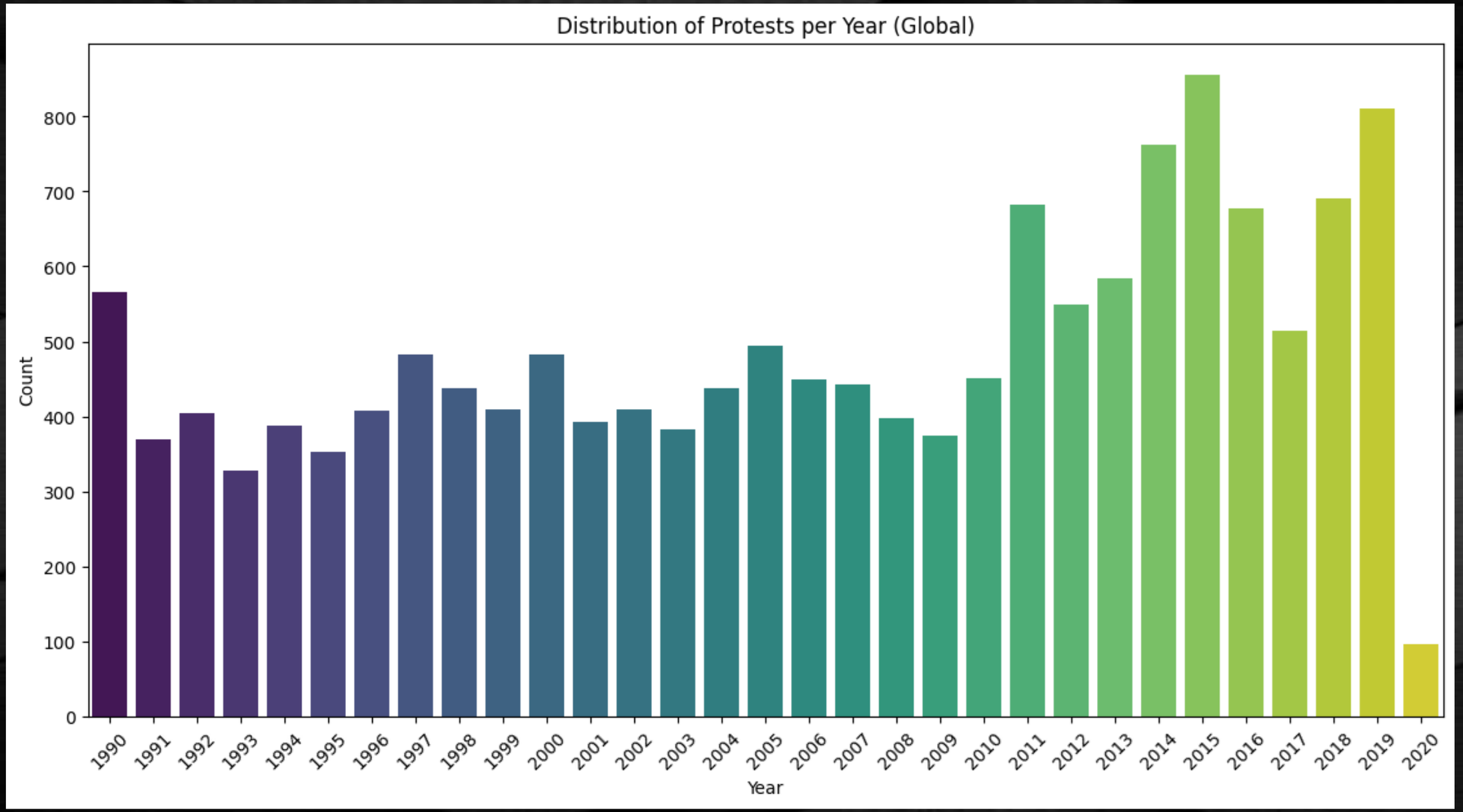
EXPLORATORY DATA ANALYSIS (EDA)

Data Insights: Provided insights into the distribution and nature of the data.

Visualizations: Created histograms, box plots, bar graphs, line plots, scatter plots and wordclouds for better understanding of the data.

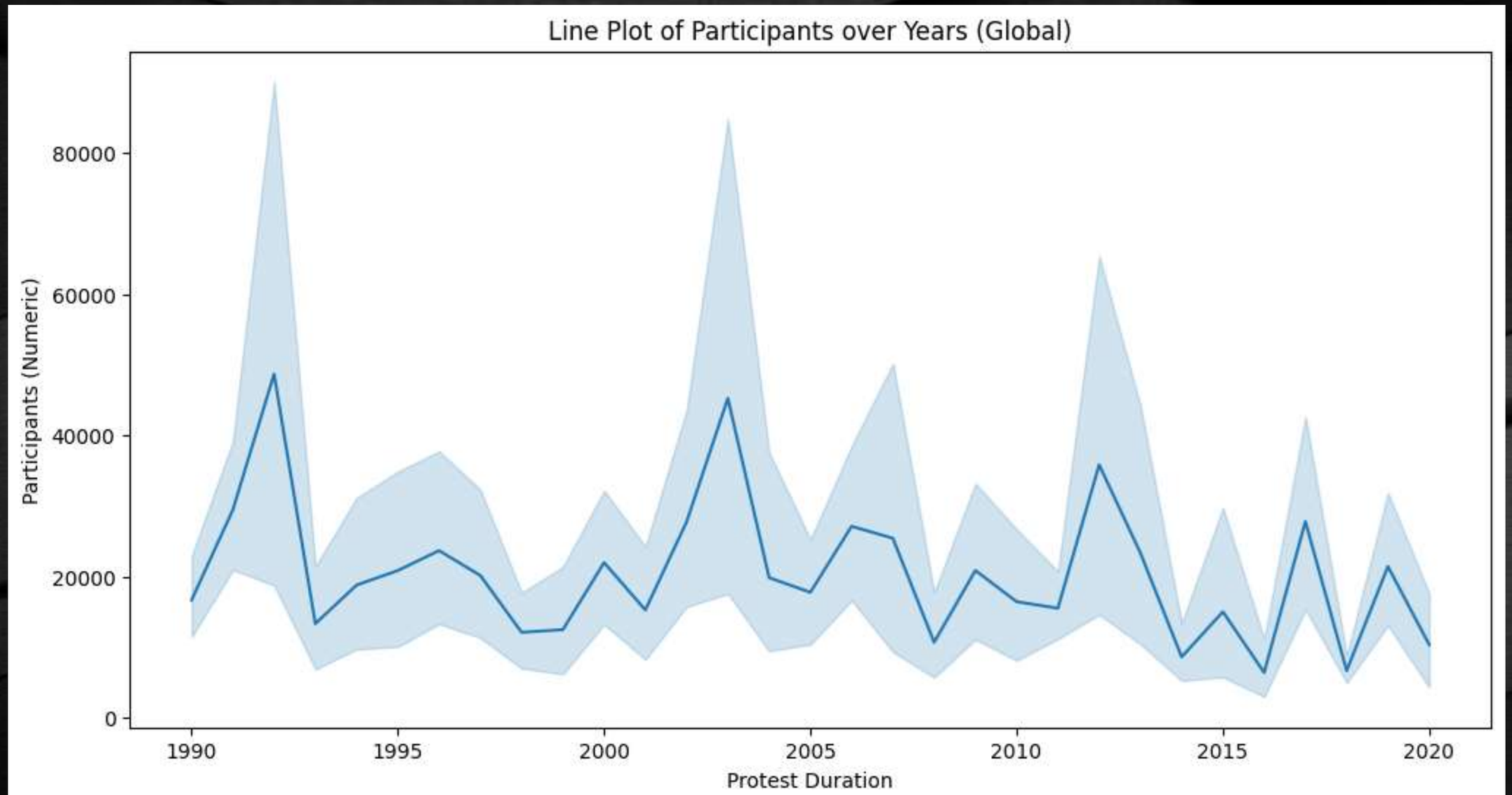


- Europe had the most recorded protests ,Africa being the second and Oceania recording the least.

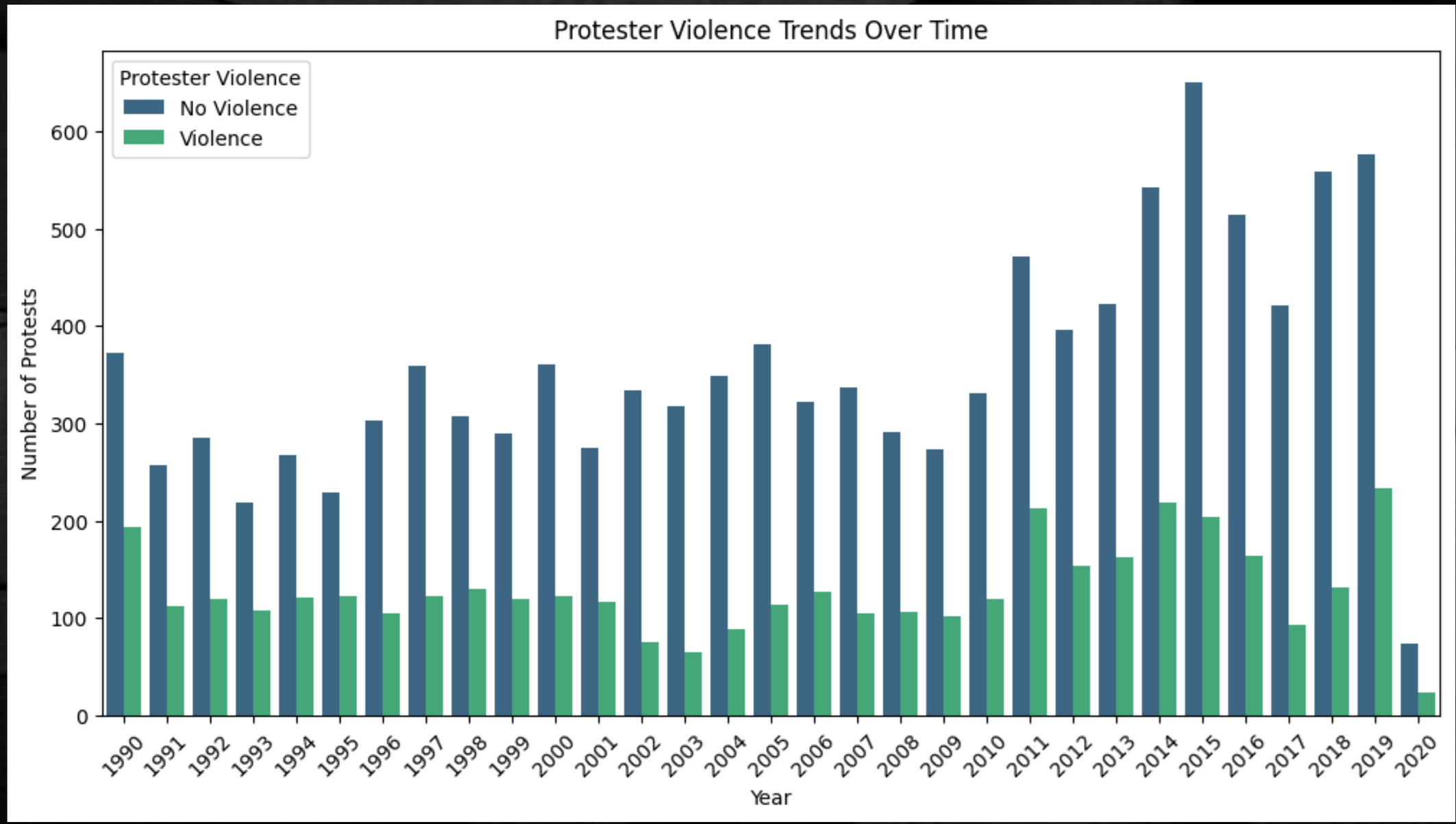


The distribution of protests took an upward turn from 2010’s up until 2020 likely influenced by the global COVID pandemic

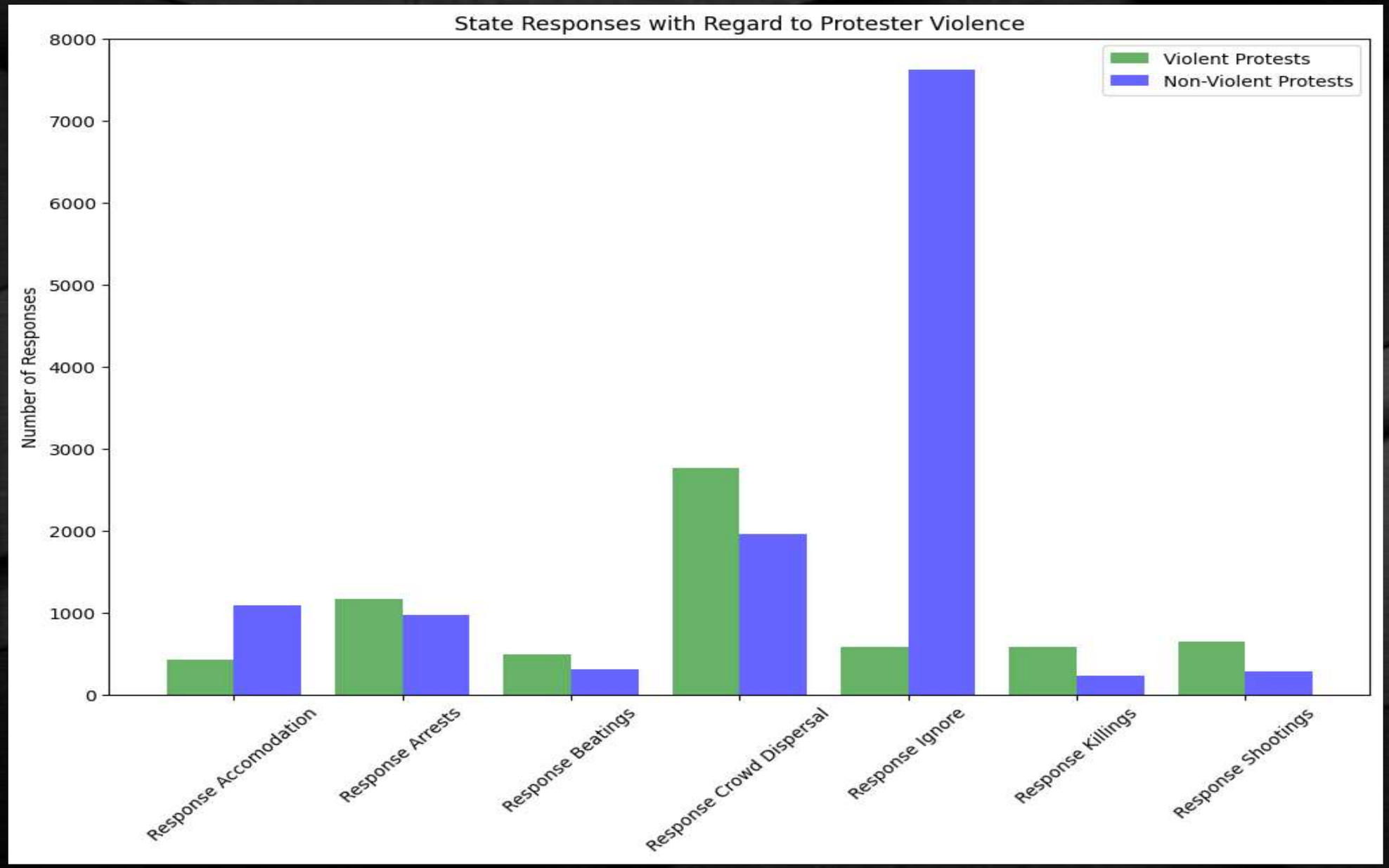
Kenya had a record high number of protests in 2015.This was also observed in other nations.



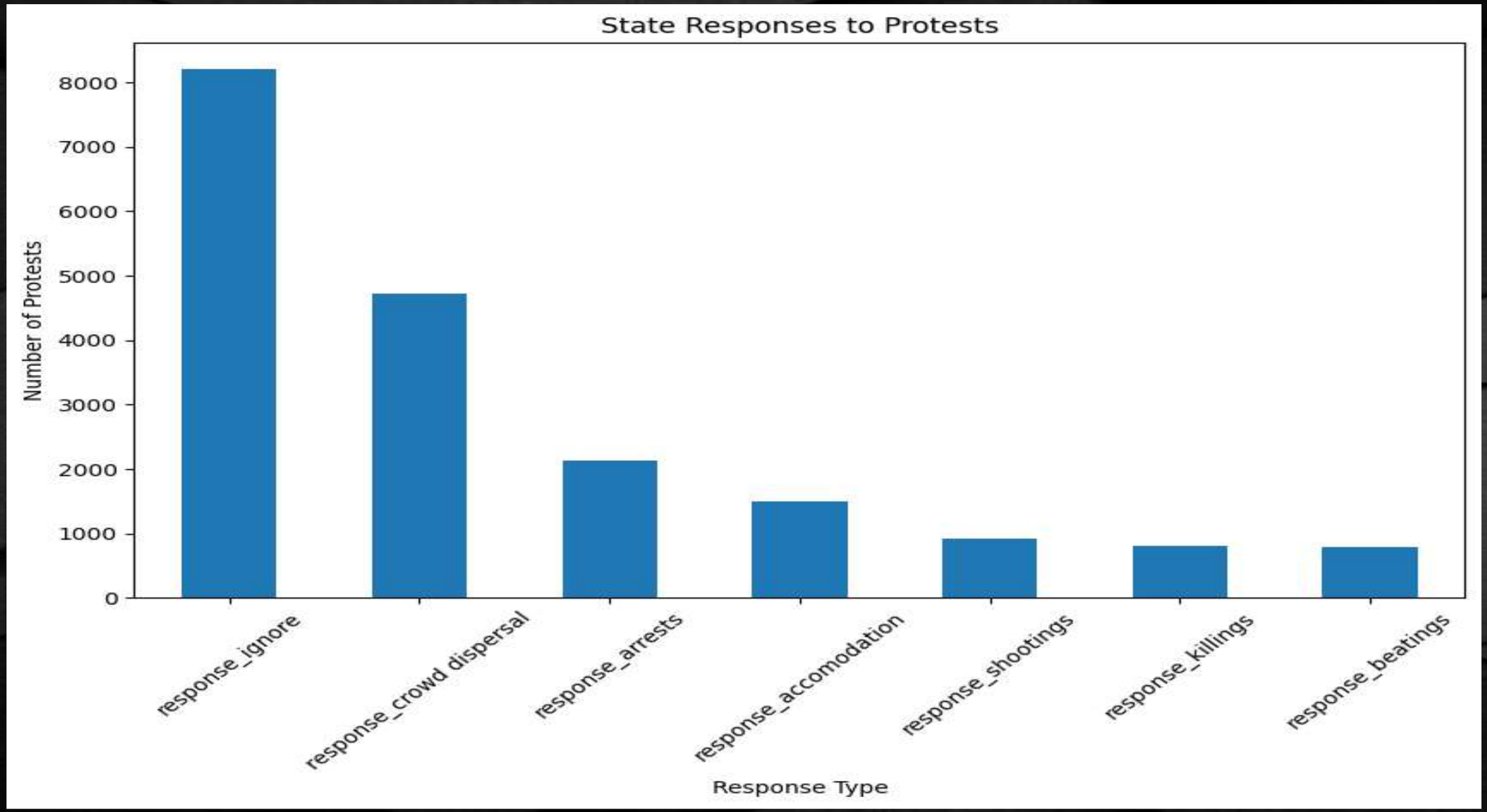
- There are noticeable spikes around the early 1990s, late 1990s, and mid-2000s, suggesting significant global protest events during these periods.
- After the mid-2000s, the number of participants in protests seems to decrease overall, with smaller spikes occurring in the 2010s.



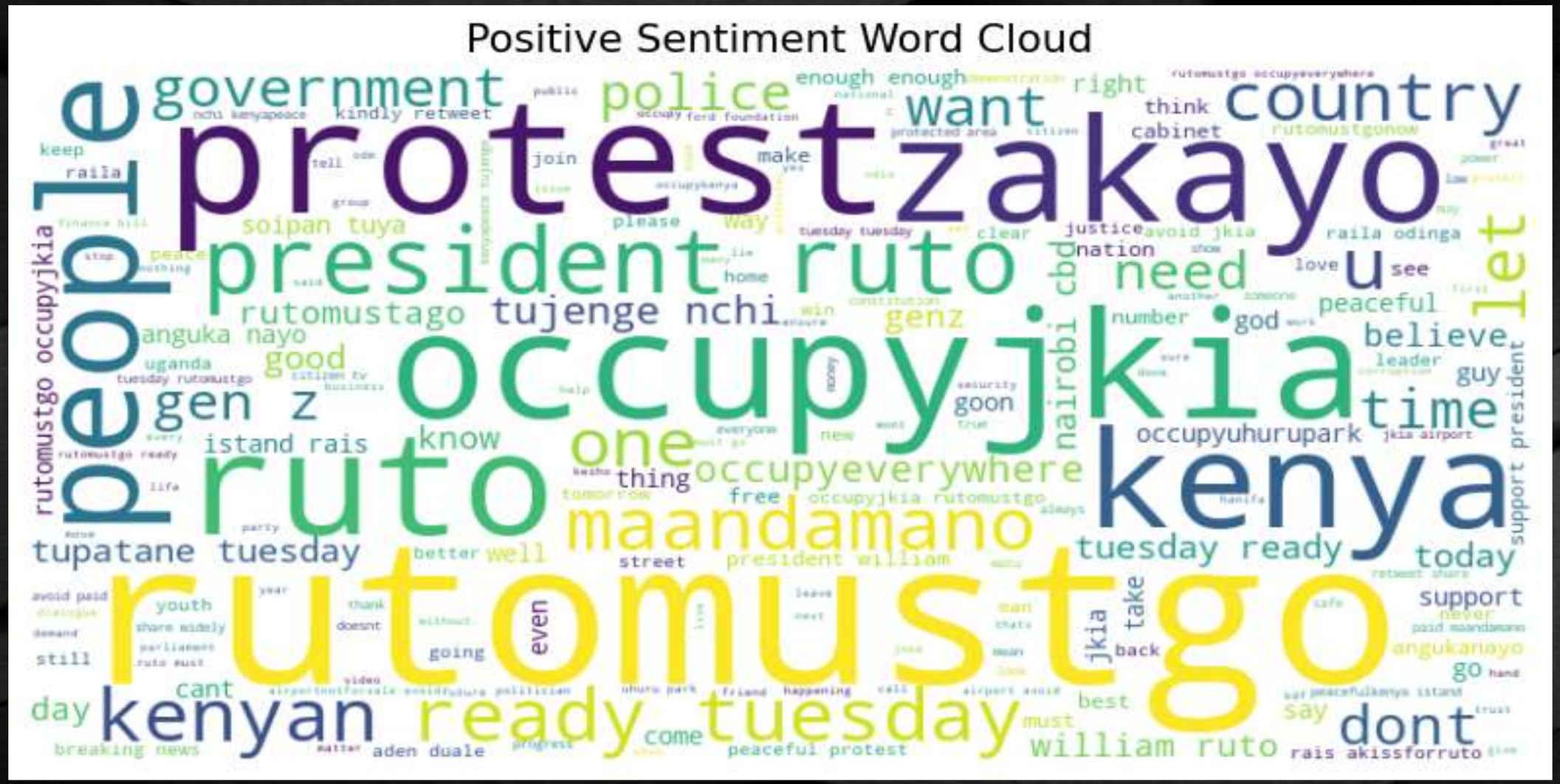
In 2015, the highest number of protests without violence was recorded, surpassing 600 incidents. Conversely, in 2019, the highest number of violent protests was observed, exceeding 200 incidents.



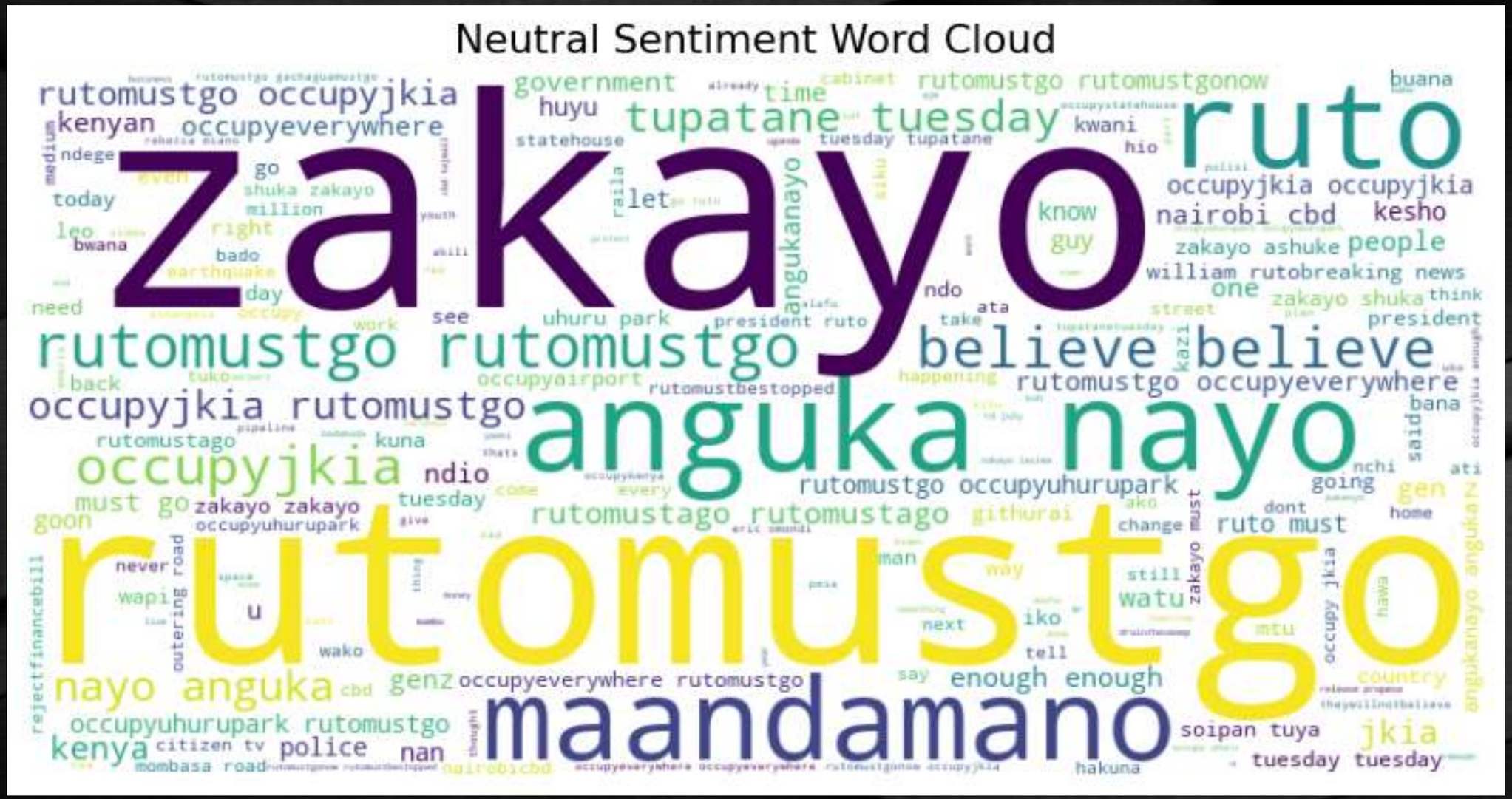
- Violent Protests: Frequent, forceful; "Severe Intervention" common
- Non-Violent Protests: Less frequent, less forceful; "Moderate Intervention" most common
- Non-Intervention: Rare overall



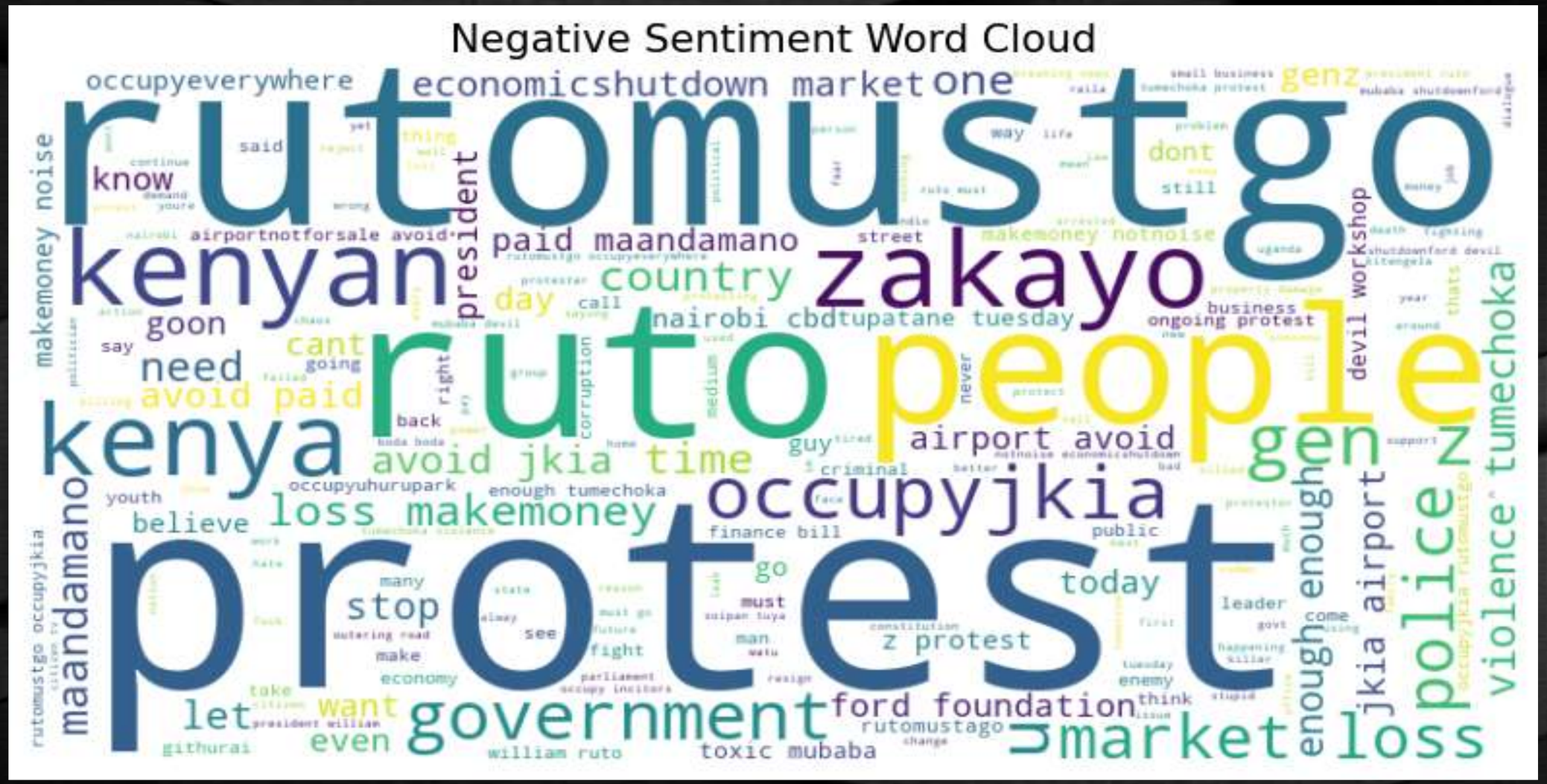
- Forceful Responses: Most common: Crowd dispersal, Arrests, then Shootings, Killings, Beatings,
- Non-Confrontational Responses: Most common: Ignoring then Accommodation



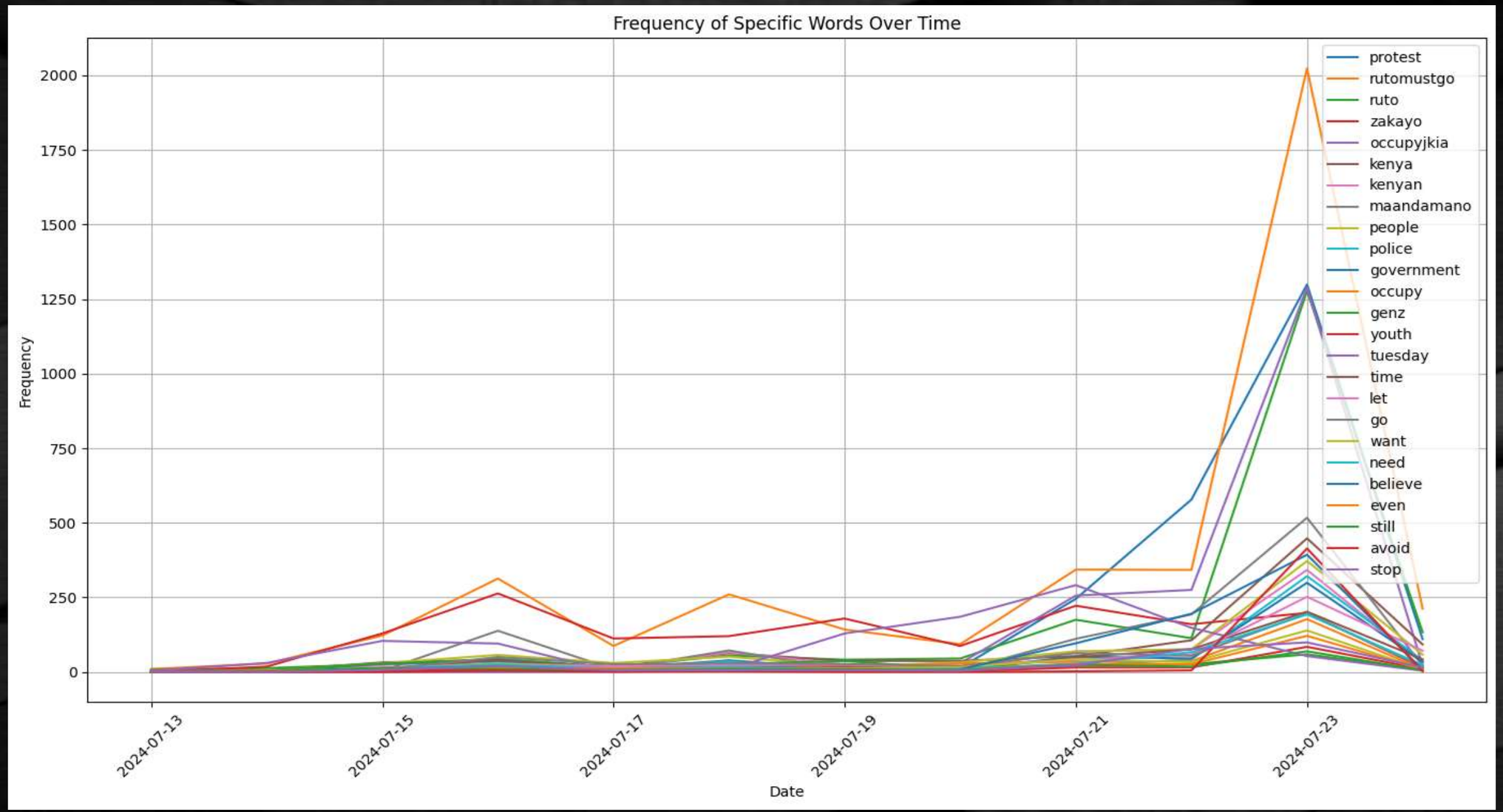
- Optimism & Unity: Positive words like "good," "peace," "justice," "believe," "support," "love.“
- Change: Terms like "protest," "occupyjkia," "rutomustgo.“
- Collective Action: Words like "people," "together," "support.”



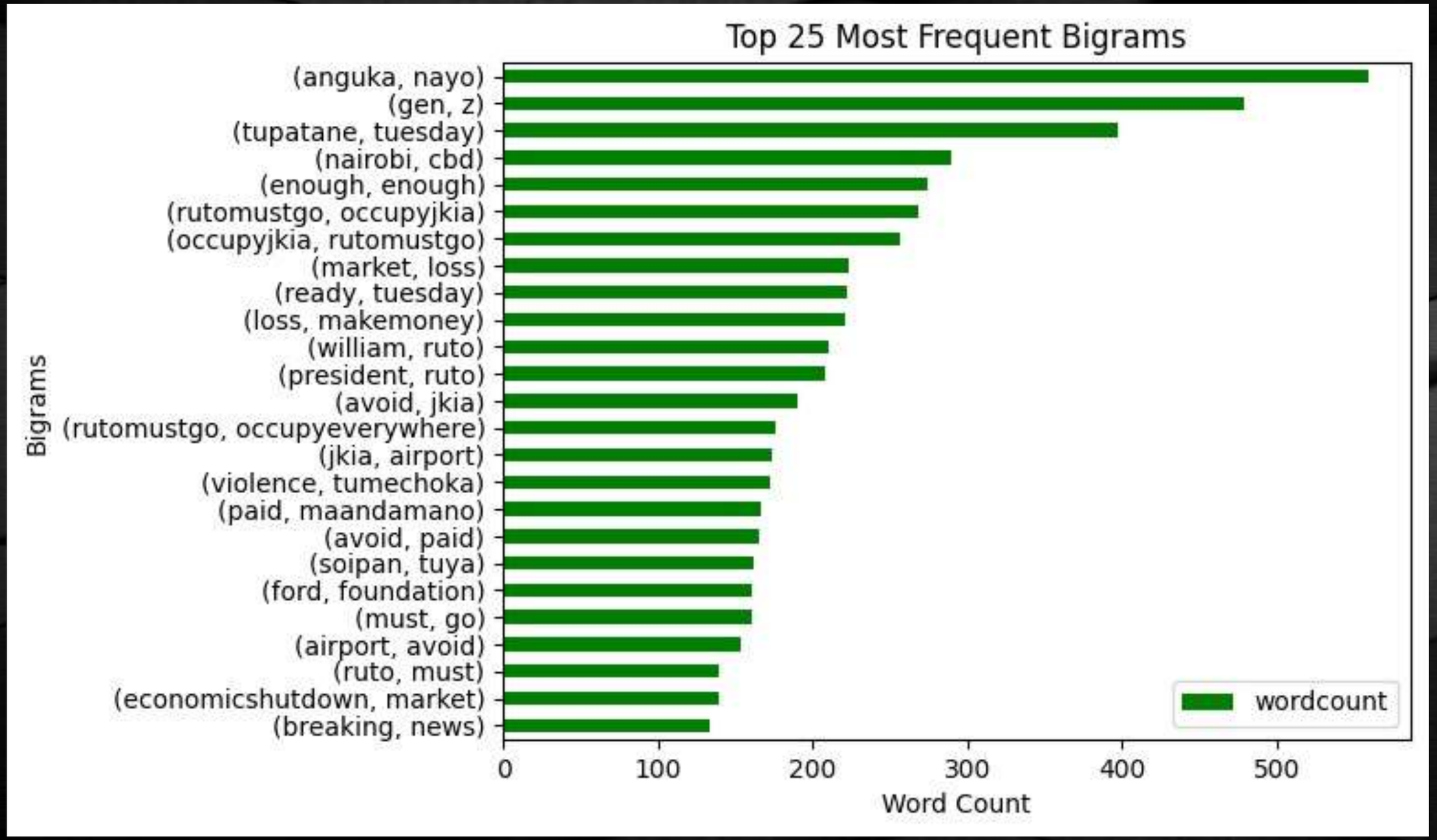
- Protest & Dissatisfaction: "Occupyjkia," "rutomustgo," "government.“
- Events & Locations: "Tuesday," "Nairobi," "Kenya.“
- Unity: "People," "together," "one.“
- Sentiment: Neutral but leaning negative.



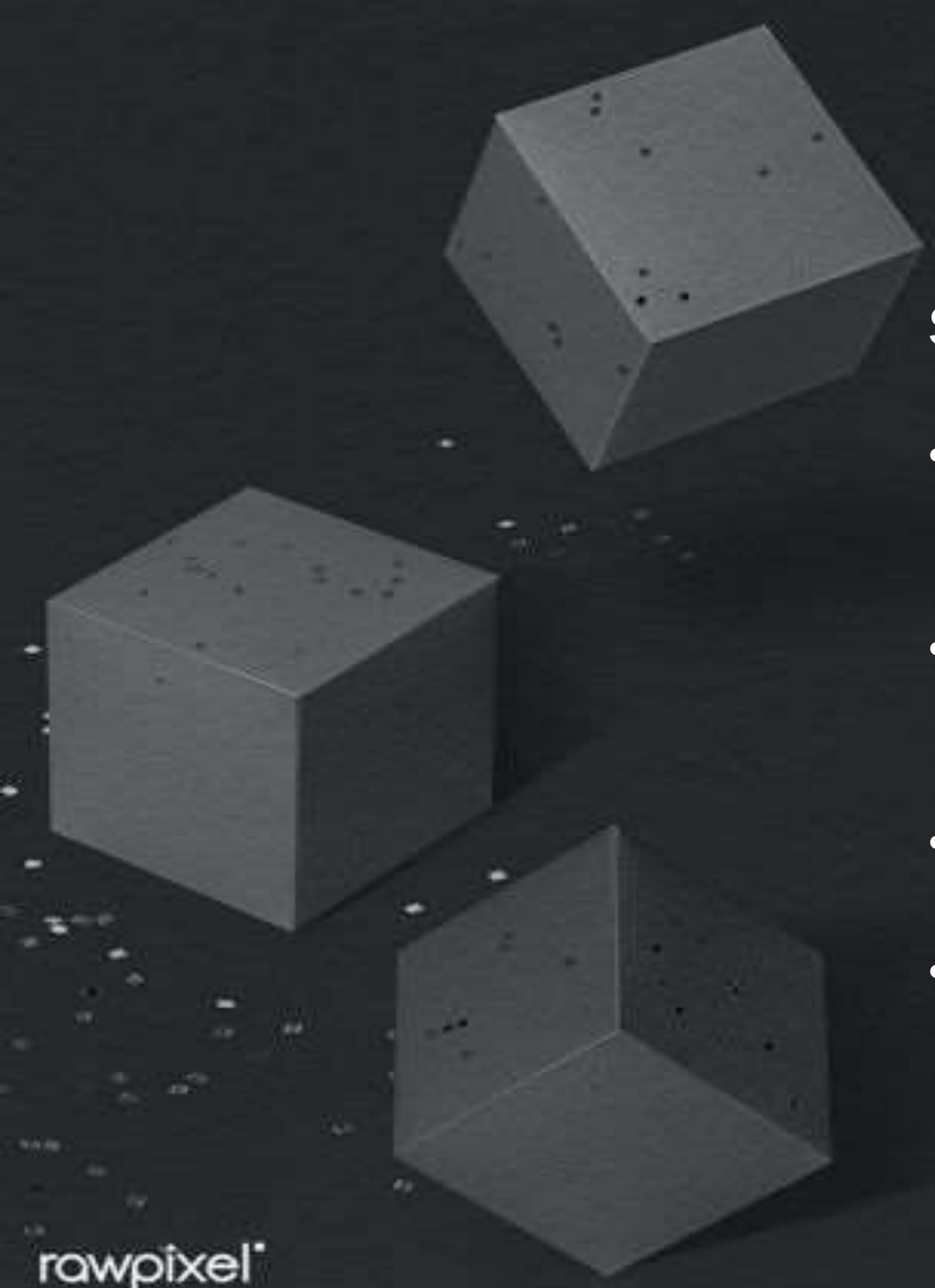
- Strong Opposition: "Maandamano," "occupyjkia," "rutomustgo," "violence“
- Economic Struggles: "market loss," "shutdown," "finance bill“
- Dissatisfaction with Government: "government," "president," "corruption“
- Calls for Mobilization: "occupyjkia," "rutomustgo"



- Trends (July 13-23, 2024): Frequent terms include "protest," "rutomustgo," "ruto," "occupyjkia.“
- Insights: Fluctuations indicate event-driven interest.
- Emerging: Rising mentions of "genz" and "youth."



Bigram: Pair of consecutive words used together in text analysis



DATA PREPARATION

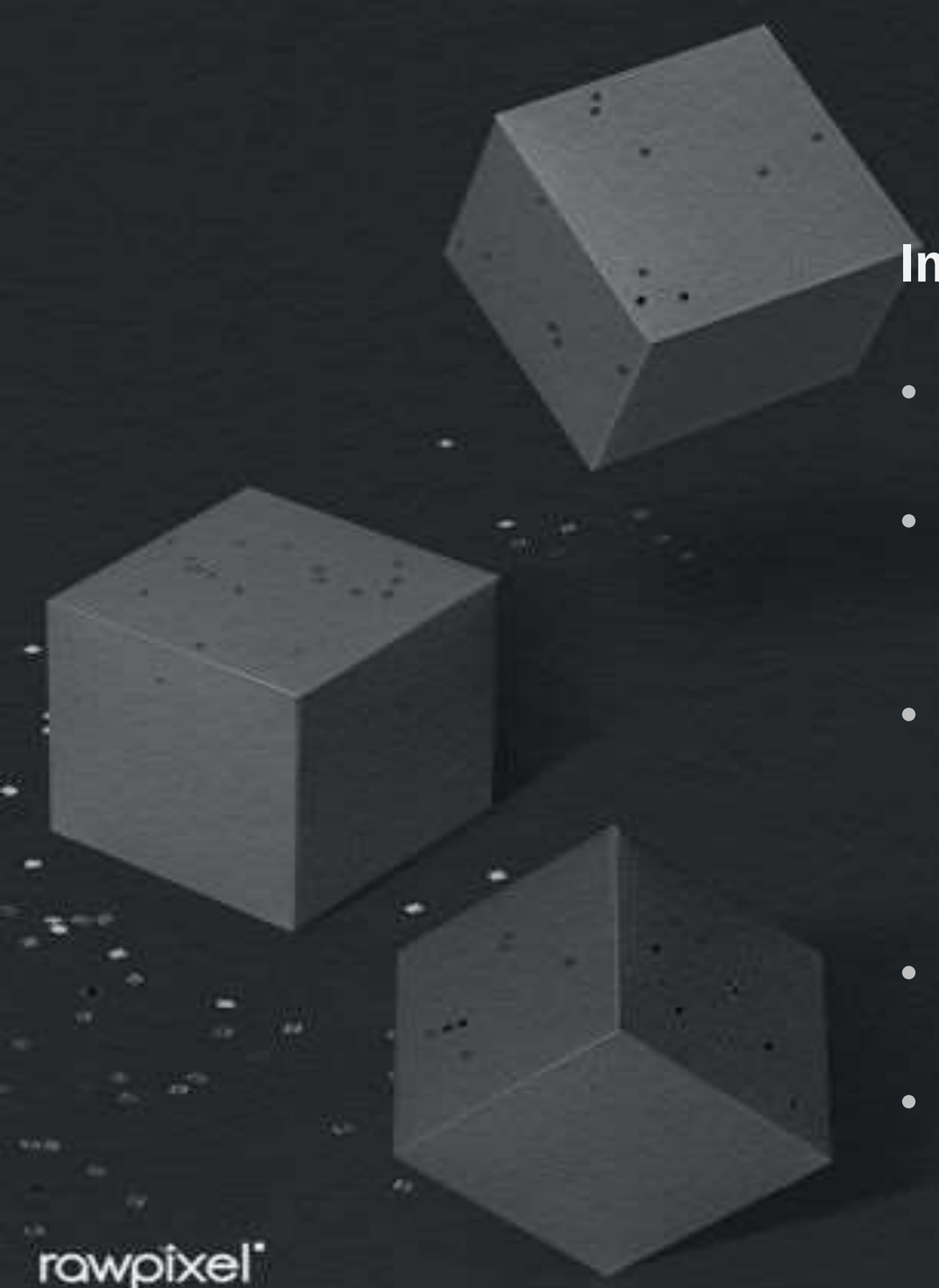
Steps:

- Cleaned data by removing extreme values in protest duration and participant numbers column
- Region analysis: Combined Americas, excluded Canada and Oceania.
- Dropped country column.
- Grouped responses into three classes: Passive/Concessive, Control Measures, and Forceful Repression to enhance analysis and modeling.

BASELINE MODEL

Logistic Regression:

- Training accuracy: 52.9%.
- Test accuracy: 52.6%.
- Performance issues: The model is not accurately identifying all instances of Control Measures, leading to some being misclassified.



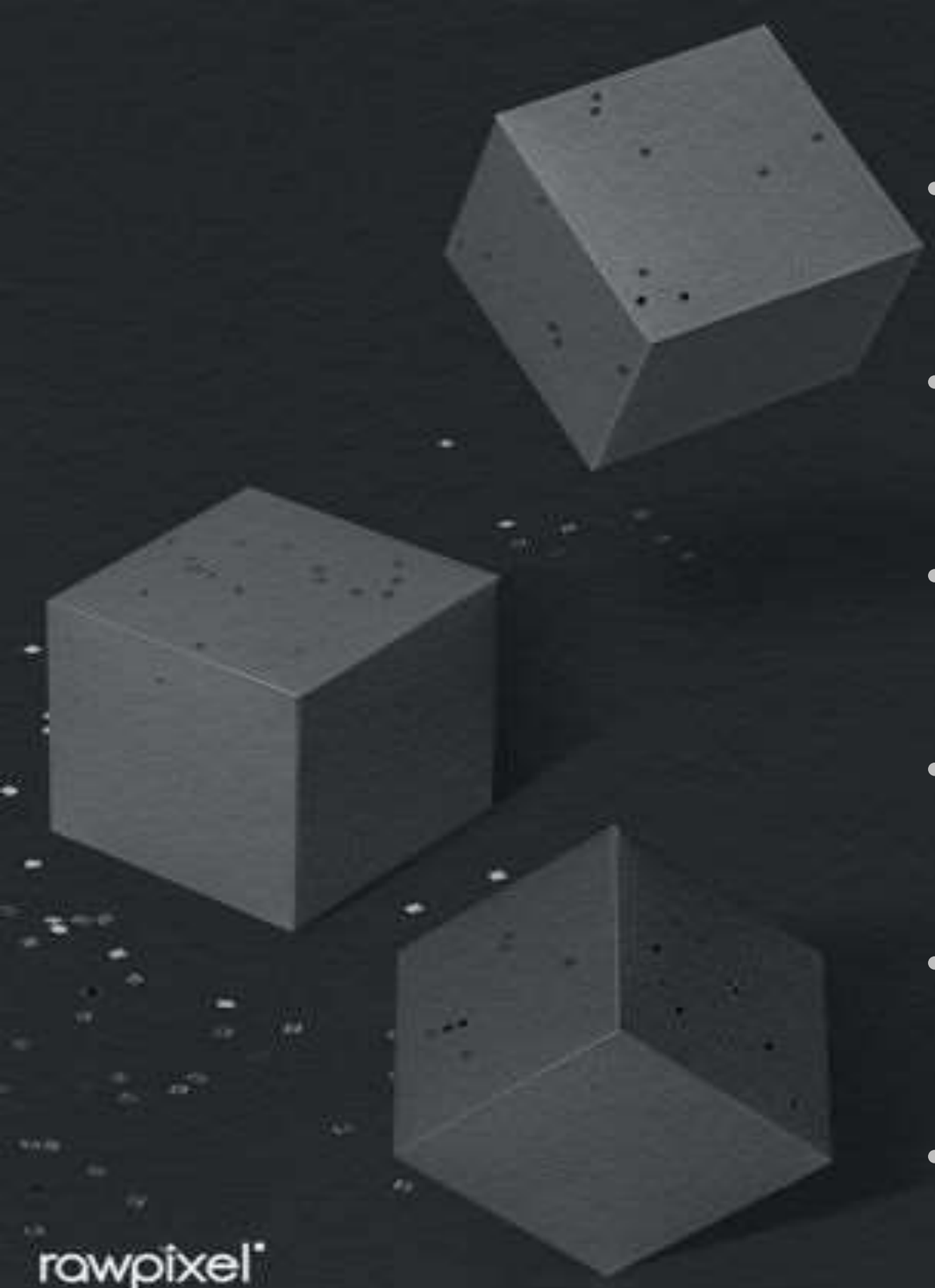
MAIN MODELS

Improvement Steps:

- Hyperparameter tuning.
- Explored XGBoost, GradientBoosting, Random Forest classifier and SVC.
- Feature engineering including LDA Topic Modelling, Sentiment Feature Extraction and Named Entity Recognition.
- Address class imbalance (SMOTE)
- Performed Linear Discriminant Analysis as a dimensionality reduction technique.

Main Models Perfomance After Tuning Vs Baseline Model:

Models	Train Accuracy	Test Accuracy
Baseline Model	0.53	0.52
Random Forest	1.0	0.79
XGBOOST	0.83	0.81
SVC	0.82	0.81



Evaluation Metrics

- Precision:
  True Positives / (True Positives + False Positives)

- Recall (Sensitivity):
  True Positives / (True Positives + False Negatives)

- F1 Score:
  Harmonic mean of precision and recall

- Confusion Matrix:
  Table of TP, TN, FP, FN

- ROC-AUC Score:
  Area under ROC curve

- Precision-Recall AUC:
  Area under precision-recall curve

BEST PERFORMING MODEL

- **Model:** XGBOOST CLASSIFIER

- **Training Accuracy:** 83%
- **Test Accuracy:** 81%

- **Classification Report:**
  Precision, Recall, F1 Score:
  Control Measures: Precision 0.76, Recall 0.74, F1 Score 0.75
  Forceful Repression: Precision 0.83, Recall 0.85, F1 Score 0.84
  Passive or Concessive: Precision 0.85, Recall 0.86, F1 Score 0.85

ROC AND PR CURVES

ROC Curves:

- Show trade-off between true positive rate (TPR) and false positive rate (FPR).
- AUC values indicate model performance:
  - Control Measures: AUC = 0.90
  - Forceful Repression: AUC = 0.96
  - Passive or Concessive: AUC = 0.96

PR Curves:

- Display precision vs. recall.
- High AUC values show effective identification:
  - Control Measures: AUC = 0.83
  - Forceful Repression: AUC = 0.91
  - Passive or Concessive: AUC = 0.92

CONFUSION MATRIX:

True Positives (Diagonal Elements): Correctly predicted instances.

- **Control Measures:** 1335
- **Forceful Repression:** 1494
- **Passive or Concessive:** 1471

False Positives and False Negatives (Off-Diagonal Elements): Instances where the model misclassified the labels.

- Misclassified as **Control Measures:** 244 (from Forceful Repression), 231 (from Passive or Concessive)
- Misclassified as **Forceful Repression:** 239 (from Control Measures), 30 (from Passive or Concessive)
- Misclassified as **Passive or Concessive:** 182 (from Control Measures), 63 (from Forceful Repression)

SUMMARY:

- The XGBoost model demonstrates robust classification performance with high accuracy and an excellent ability to distinguish between classes.

- The high AUC values across ROC and PR curves indicate the model's strong discriminative power, and the confusion matrix reflects a high number of correct predictions with relatively few misclassifications.

- This analysis underscores the model's effectiveness in classifying instances across the given categories

RECOMMENDATIONS:

- Further Model Tuning: Further hyperparameter tuning and explore other algorithms like ensemble methods to potentially boost performance further. Implement advanced techniques such as ensemble learning (e.g., stacking models) to leverage the strengths of multiple algorithms.

- Regular Model Evaluation: Continuously evaluate the model with new data to ensure its performance remains robust over time. Regularly update the model with recent data to maintain its accuracy and relevance.

- Feedback Loop: Establish a feedback loop with stakeholders to incorporate their insights and concerns into model improvements and data collection efforts.

CONCLUSION:

- High Model Performance: The XGBoost model showed excellent accuracy and effectiveness in distinguishing state responses, particularly excelling in detecting Forceful Repression and Passive or Concessive responses.
- Insightful Sentiment Analysis: Analyzing tweets provided valuable insights into public sentiment and protest trends, enriching the understanding of protest dynamics.

LIMITATIONS:

- Class Imbalance: The dataset's class imbalance, especially for Control Measures, impacted the model's prediction accuracy and led to higher misclassification rates.

- Data Quality and Features: Variability in data quality and limited features constrained the model's ability to fully capture the complexity of protest dynamics and state responses

- Lack of Real-Time Data: The model relies on historical data, which may not reflect current protest dynamics and state responses accurately.

DEPLOYMENT

These are our deployment sites:

- **Protest Outcome Prediction Model** https://protests.streamlit.app/

- **Mass Mobilization Analytics** https://mass-protests.streamlit.app/

thank you