Sentiment Analysis of Apple Tweets

GROUP 8

This presentation outlines a machine learning project focused on analyzing public sentiment towards Apple and Google on Twitter.

Business Understanding: The Importance of Social Sentiment

Overview

Google and Apple's brand reputation is significantly influenced by social media discourse. Real-time analysis of public sentiment on platforms like Twitter is crucial for informed decision-making across marketing, public relations, and investment strategies. Understanding the public pulse allows for proactive issue management and capitalizing on positive trends.

Problem

Manually processing thousands of tweets to gauge sentiment is inefficient, resource-intensive, and prone to human error. Automating this process through machine learning provides a scalable, consistent, and rapid solution.

Objectives

1

To pre process the tweet data using Natural Language Processing techniques

2

To build a **machine learning classifier** that accurately predicts the sentiment of tweets.

3

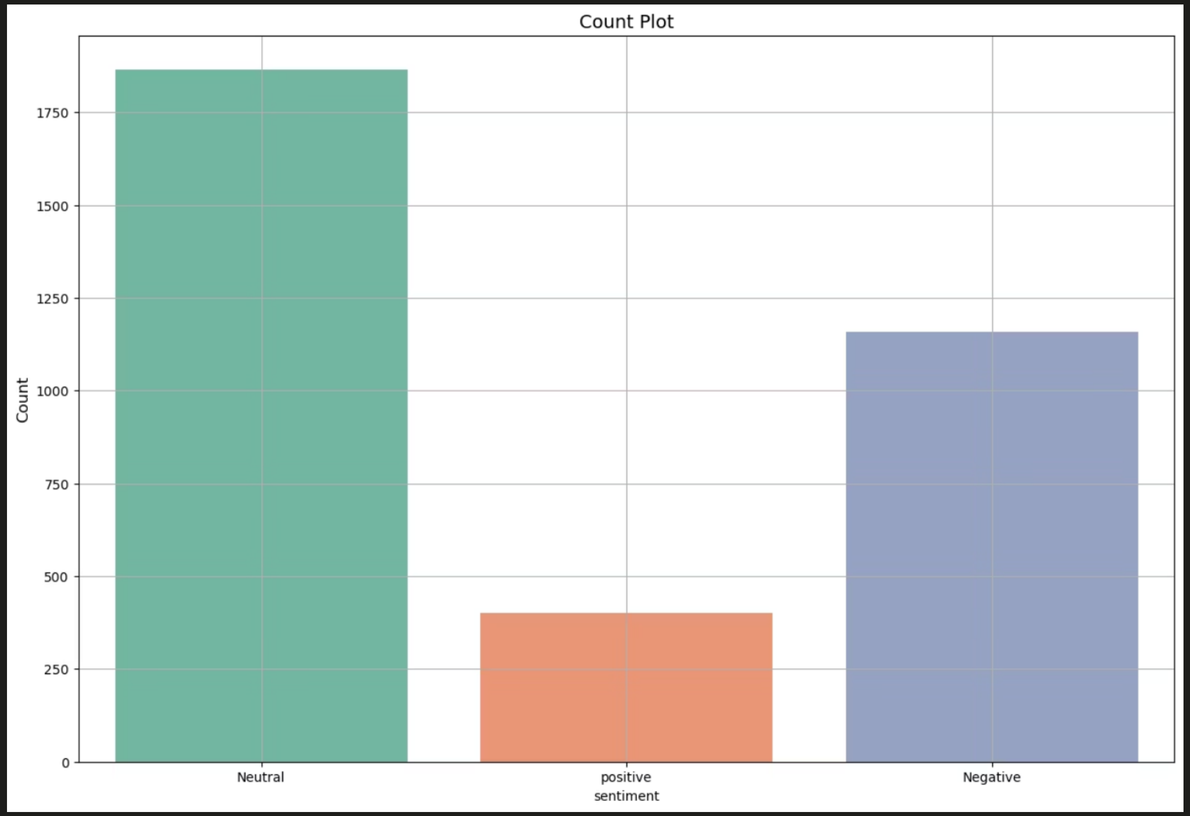
To evaluate the performance of different classifiers using appropriate **classification metrics** such as F1-score and roc-auc score

4

To Deploy the best performing models

Data Understanding: Dataset Overview

- **Size:** The dataset comprises 3,886 unique tweets, providing a substantial basis for model training and validation.
- **Labels:** Each tweet is annotated with a sentiment label: positive, neutral, or negative, crucial for supervised learning.
- **Additional Fields:** Beyond the core text and sentiment, the dataset includes metadata such as 'confidence' scores, 'tweet text', and 'date', which can be leveraged for deeper insights or data integrity checks.
- **Origin:** The data we used was from Kaggle(Apple-sentiment-Analysis-csv file) and the was collected via crowdsourcing and had manual labeling applied, contributing to the quality and relevance of the sentiment labels.



Data Preparation: Cleaning and Feature Engineering

Cleaning Steps

- filtered dataset to retain only relevant rows and columns.
- Renamed and casted columns for consistency.
- Removed nulls and duplicates.
- Cleaned text: removed punctuation, links, emojis, and stopwords.
- Applied tokenization and lemmatization using NLTK.

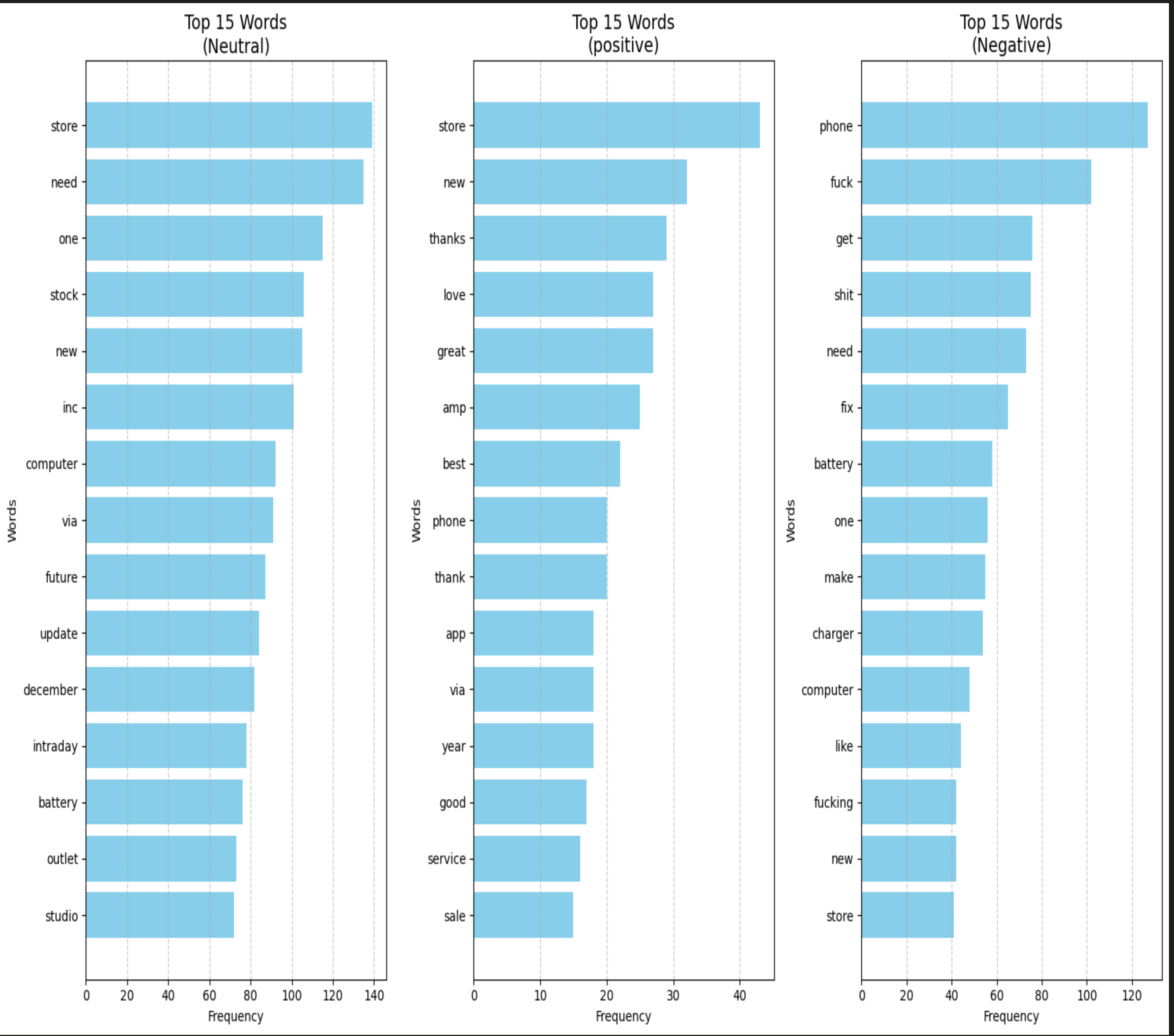
Feature Engineering

Engineered new features:

- Tweet length
- Word/sentence counts
- Lexical diversity

Augmentations:

- SMOTE applied to handle class imbalance.
- TF-IDF vectorization for model input.
- Did text augmentation for the minority class



- In our data the top word per sentiment were: store, new, thanks for positive. Need, new ,stock for neutral and fix, make phone and suck for negative sentiments

Modeling: Candidate Selection & Refinement

1

Initial Candidates

A broad range of classifiers were initially considered for their diverse algorithmic approaches:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Trees
- Random Forest Classifier
- XGBoost

2

Top Performers

After initial cross-validation and benchmarking, the following models emerged as top-tier candidates for further refinement:

- Logistic Regression (chosen as baseline)
- Random Forest Classifier
- XGBoost

3

Model Improvements

To enhance performance and address specific challenges, each top model underwent significant improvements:

- **SMOTE Integration**
- **Augmented Features**
- **Hyperparameter Tuning**

This systematic approach ensures that the selected models are not only effective but also robust and tailored to the unique characteristics of the tweet sentiment data.

Evaluation: Metrics & Performance

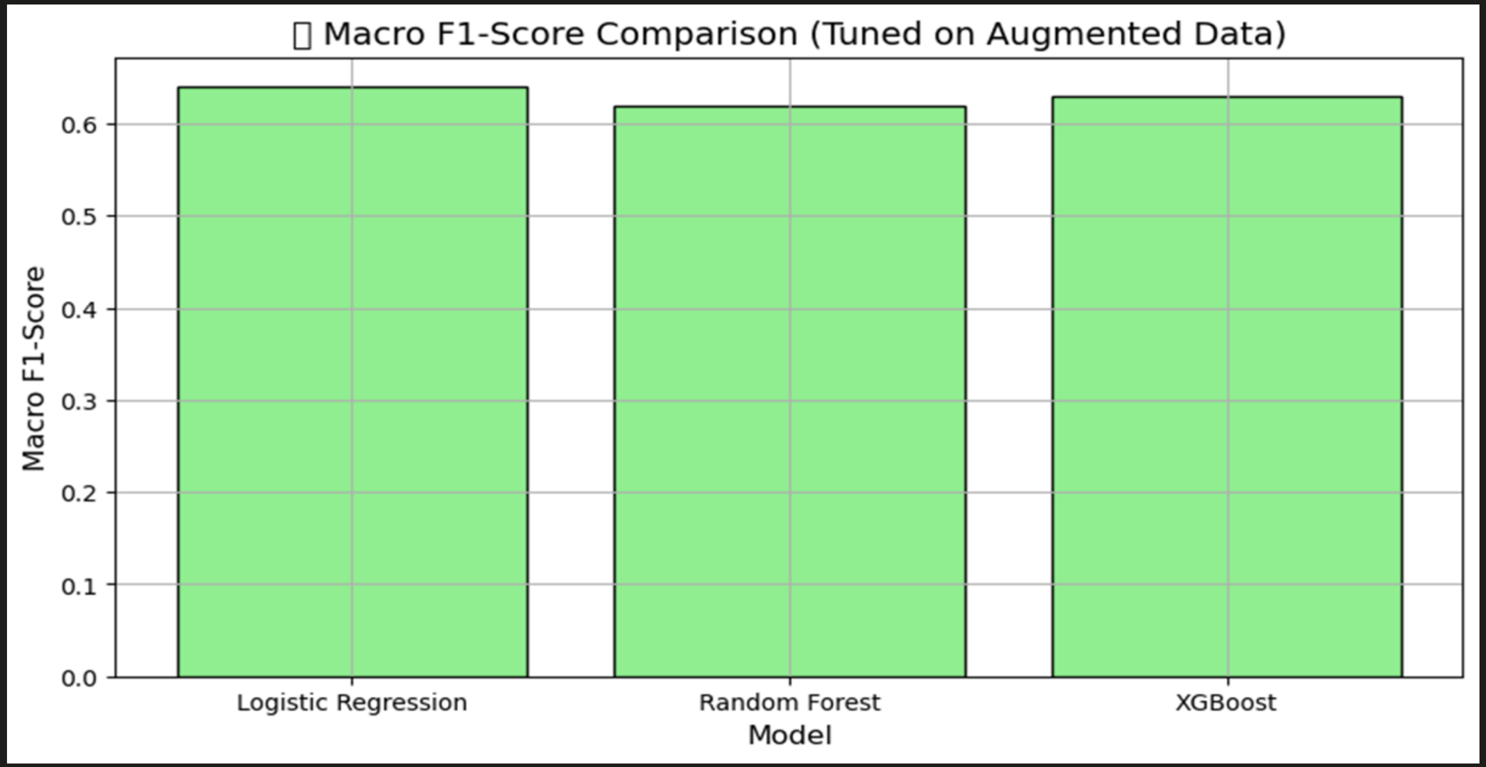
The performance of all tuned models was rigorously evaluated using a comprehensive suite of classification metrics.

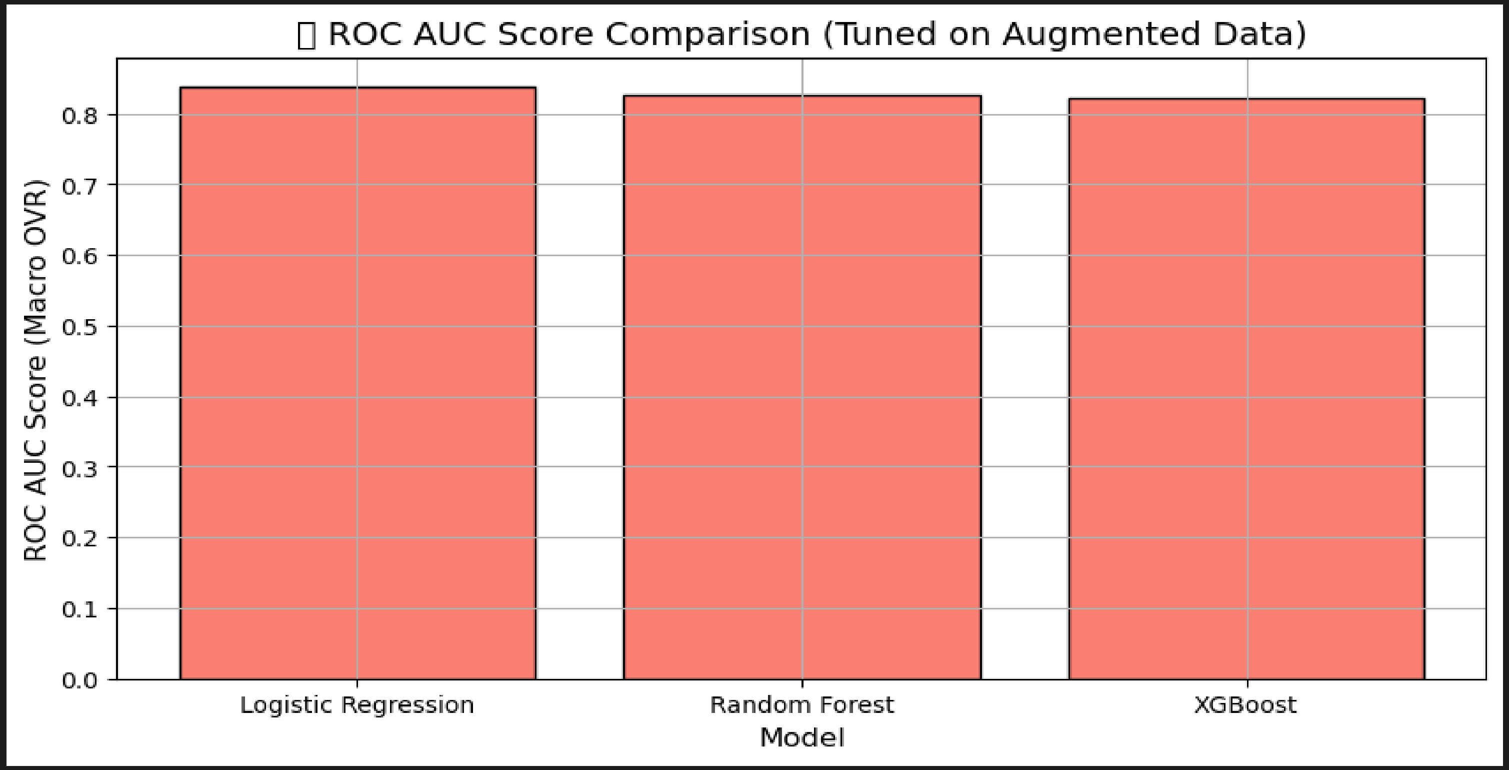
Key Metrics:

- Precision
- Recall
- F1 Score
- ROC AUC Score

Observations & Decision:

- **Logistic Regression Selected:** Logistic Regression emerged as the most suitable. It portrayed:
  - **Highest ROC AUC:**
  - **Competitive Recall**
  - **Interpretability & Stability**





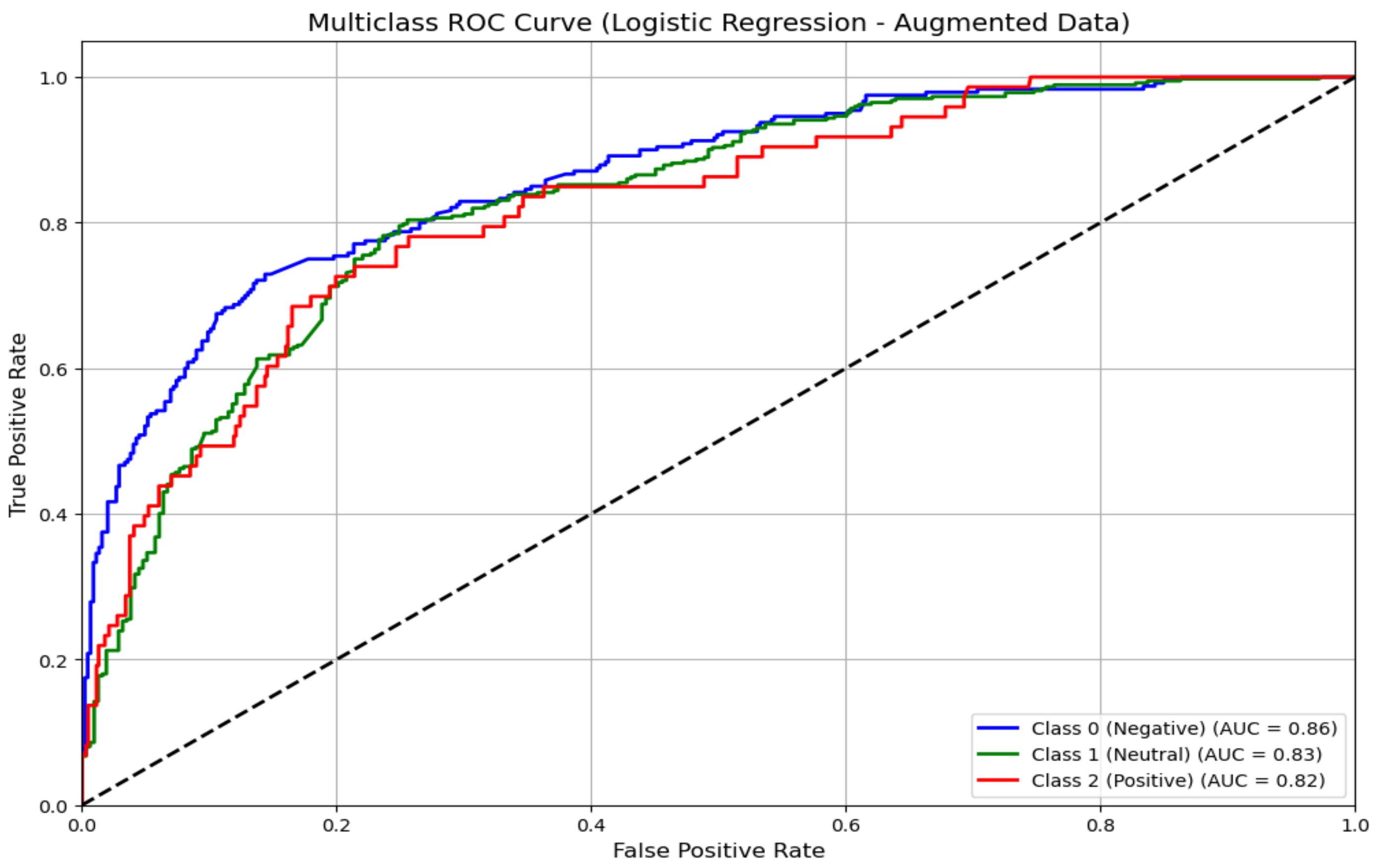
Evaluation: Final Model Analysis

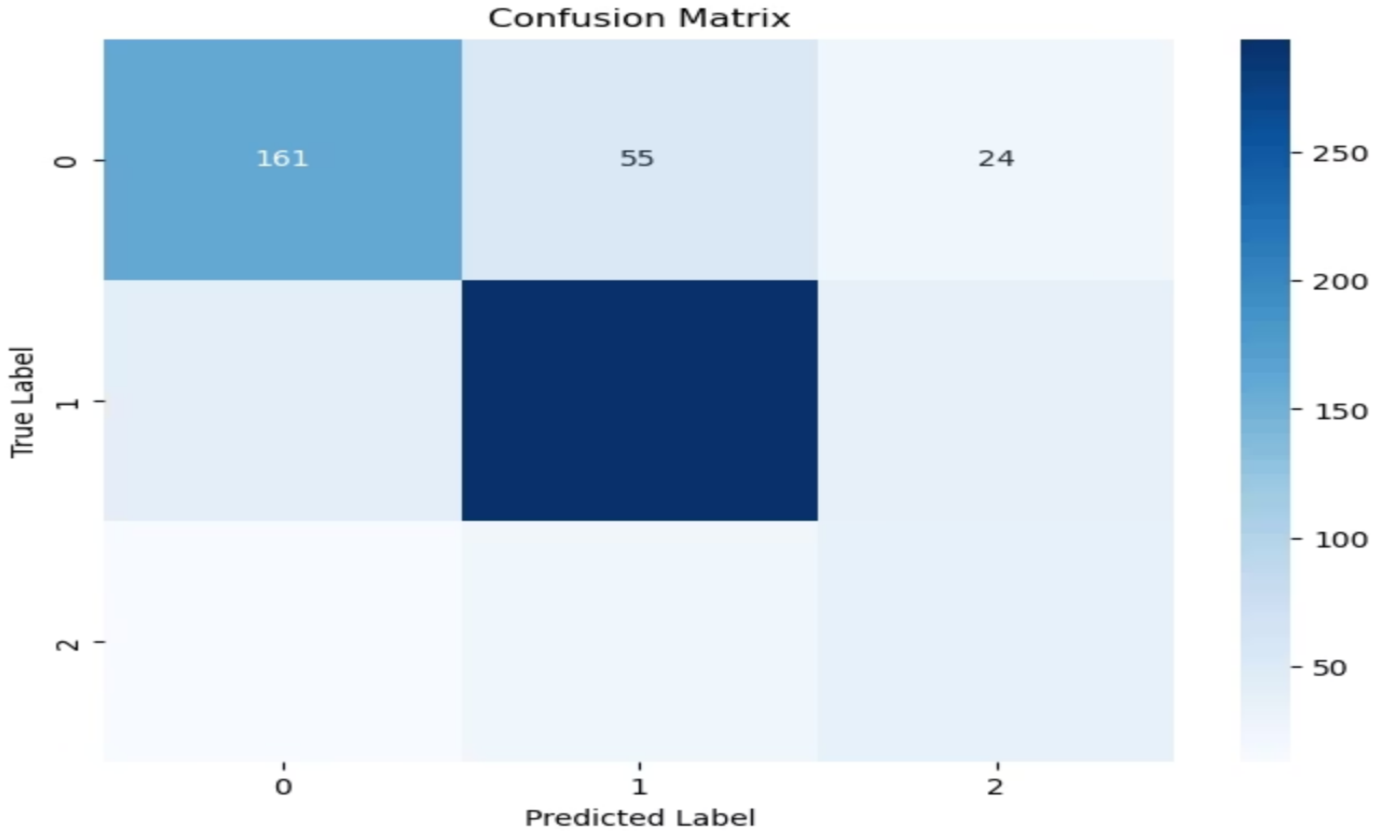
A deeper, multi-faceted evaluation of the selected Tuned Logistic Regression model was conducted to thoroughly assess its performance, generalizability, and interpretability.

Analysis Components:

- **Classification Report**
- **Confusion Matrix**
- **ROC Curve**
- **Model Interpretability with LIME**
- **Feature Importance**

❖ The trained model was saved using Joblib to facilitate quick deployment and prevent redundant retraining, ensuring efficiency for future use.





Recommendations And Future Work

Based on our analysis, we recommend the following strategic deployment and future development steps:

1. **Deploy Tuned Logistic Regression**
2. **Regularly retrain the model with updated tweets**
3. **Explore Advanced Models for future gains**
4. **Use sentiment analysis results to drive actionable outcomes i.e:**
   - ❖ **Monitor brand sentiment (e.g., Apple) in real-time.**

Challenges Faced and Assumptions

- **Assumption of Text Representation**
- **Quality and Bias in Augmented Data**.
- **Assumption of Data Stationarity**.
- **Subjectivity in Sentiment Labels**

Thank you!
Any Questions?

Group 8:

I. Kevin Kibet
II. Vincent Ngochoch
III. Chris Gitonga