NATURAL LANGUAGE PROCESSING (NLP) GROUP 7

- MEMBERS:
- 1.TRACY GWEHONA (BUSINESS UNDERSTANDING & DATA CLEANING)
- 2. BOSCO MUKARA (DATA CLEANING & EDA)
- 3.EDWIN MAINA (DATA PREPROCESSING & DEPLOYMENT)
- 4. STEPHEN MUNGAI (MODELLING)
- 5.KELVIN MWANGI (SCRUM MASTER, README.md & PRESENTATION)

OVERVIEW

- This project aims to analyze Twitter sentiment about Apple and Google products using Natural Language Processing (NLP). The dataset contains tweets labeled as positive, negative or neutral. By building a sentiment analysis model, we aim to categorize the sentiment of tweets accurately and gain insights into public perception of these tech giants' products.

BUSINESS PROBLEM

- Understanding customer sentiment is critical for businesses to gauge public opinion and improve products or services. For Apple and Google, analyzing Twitter sentiment can provide actionable insights to enhance user satisfaction and market strategies.

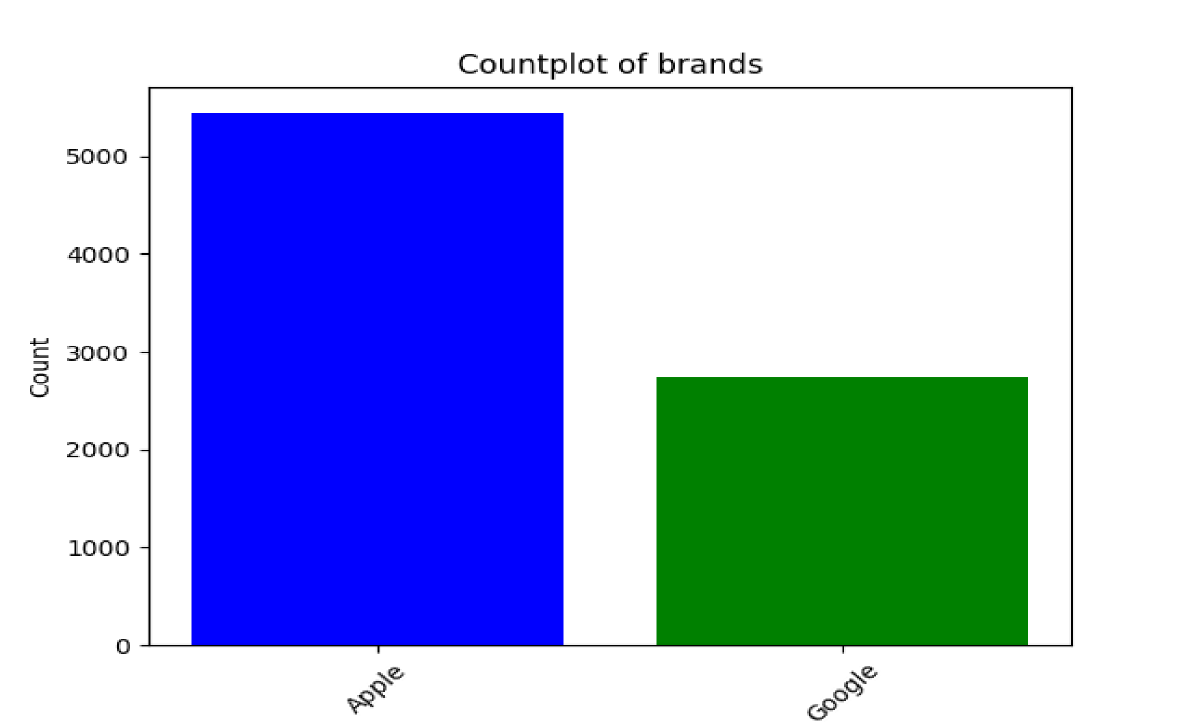
OBJECTIVES

- Build a model to classify the sentiment of tweets into positive, negative, or neutral categories. Evaluate model performance using suitable metrics. Provide insights and recommendations based on the analysis results.

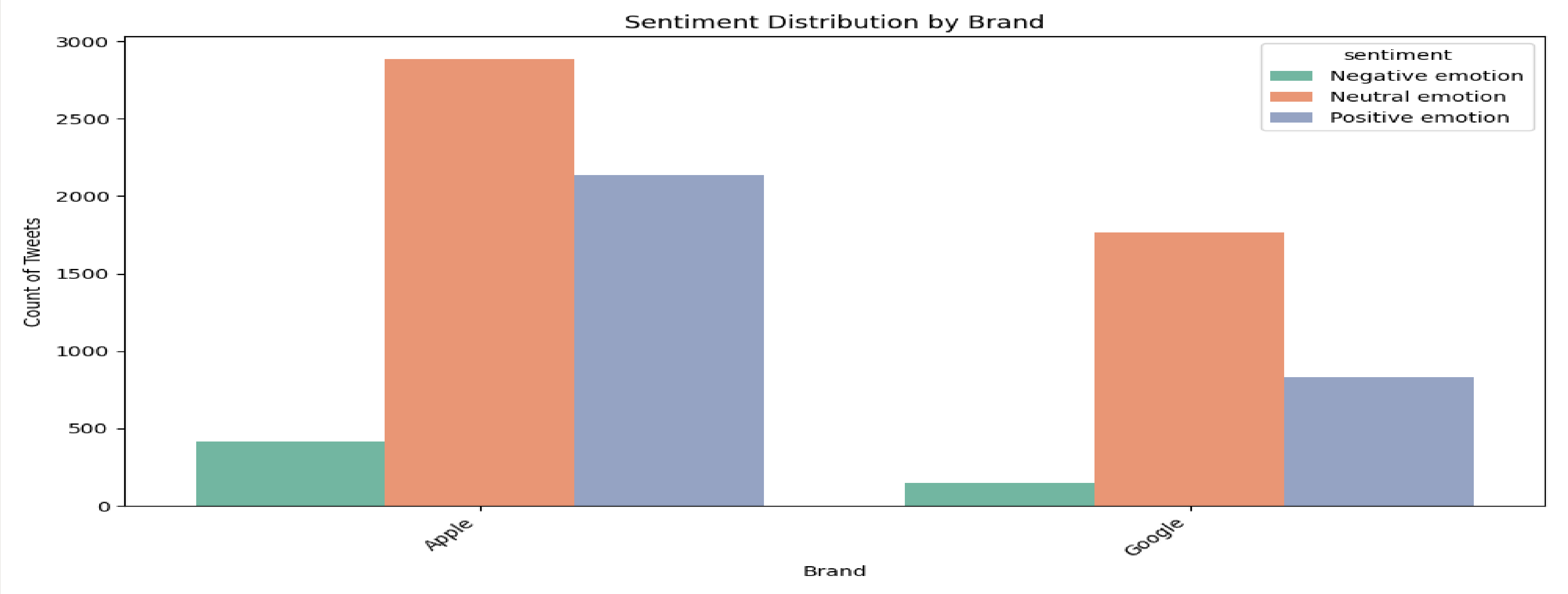
DATA UNDERSTANDING AND INSPECTION

- Here, we intend to comprehensively explore and analyze our dataset to gain insights into its structure, content, and quality.
- The dataset contains tweets labeled as positive, negative, or neutral
- Note : For the presentation we'll focus on the visualizations only.

VISUALIZATIONS (COUNT PLOT)



SENTIMENT DISTRIBUTION BY BRAND



MODELS UTILIZED

- Logistic Regression
- Random Forest
- Multinomial Naive Bayes
- Support Vector Machine
- The Logistic Regression model with TF-IDF Vectorizer and SMOTE emerged as the best-performing model

RECOMMENDATIONS

- Adopt the Best Performing Model The Logistic Regression model with TF-IDF Vectorizer and SMOTE is the most suitable for sentiment analysis due to its strong performance metrics:
- F1 Score: 0.6512 (indicating good balance between precision and recall).
- ROC AUC: 0.7519 (showing strong ability to distinguish between sentiments).
- Consistency: Low standard deviation of 0.0064, ensuring reliable predictions across data samples.

CONCLUSIONS

- The Logistic model emerging the best among the models should be adopted for use and deployment

THANKYOU

Github : https://github.com/mwangikelvin201/GROUP7-Phase4