Sentiment Analysis for Product Insights

This project **aims to develop a Natural Language Processing (NLP) model to analyze sentiment in Tweets related to Apple and Google products. By classifying the sentiment of these Tweets as positive, negative, or neutral**, the model will provide valuable insights into public perception, aiding businesses in marketing strategies and product development.

By Group 4

Business Problem

1 **Tracking Customer Sentiment**

In an era dominated by social media, brands must continuously track customer sentiments expressed online. Twitter has become a critical platform where users voice their opinions about products and brands.

2 **Manual Analysis Challenges**

The vast volume and rapid pace of tweets make it impractical for businesses to manually analyze these opinions for insights. An NLP model is needed to automatically classify the sentiment of tweets and determine which brand or product is the target.

Objectives

1. To develop a binary classification model that classifies tweets as either positive or negative.

2. Expand the binary classifier to include neutral sentiments, creating a multiclass classifier using various models

3. The goal is to build an NLP model that can accurately and efficiently classify sentiments, identify brands/products, and handle ambiguity in tweets.

Methodology Overview

1. Data Understanding

□ Analyze the dataset, which includes columns such as *tweet_text, emotion_in_tweet_is_directed_at, and is_there_an_emotion_directed_at_a_brand_or_product.*

□ Address any anomalies like missing values and duplicates.

2. Data Preparation

□ Remove duplicate entries.

□ Populate missing values in th emotion_in_tweet_is_directed_at column with "none.“

□ Apply text preprocessing techniques: tokenization, lowercasing, stopword removal, and lemmatizati

3. Modeling

□ Utilize libraries like NLTK (for tokenization, stopword removal, lemmatization), sklearn's CountVectorizer (for vectorization), and pandas (for data handling).

□ Build a logistic regression model for binary classification (positive/negative sentiment), aiming for 70% accuracy.

□ Expand the model to a multiclass classifier to include neutral sentiments.

4. Evaluation

□ Use accuracy as the primary evaluation metric, assessing the model’s ability to classify sentiments correctly.

□ Address potential limitations such as missing values and data quality issues.

Dataset Overview

Dataset Size

The dataset from CrowdFlower includes over 9,000 tweets that have been evaluated for sentiment (positive, negative, or no emotion) and tagged with the associated brand or product.

Sentiment Distribution

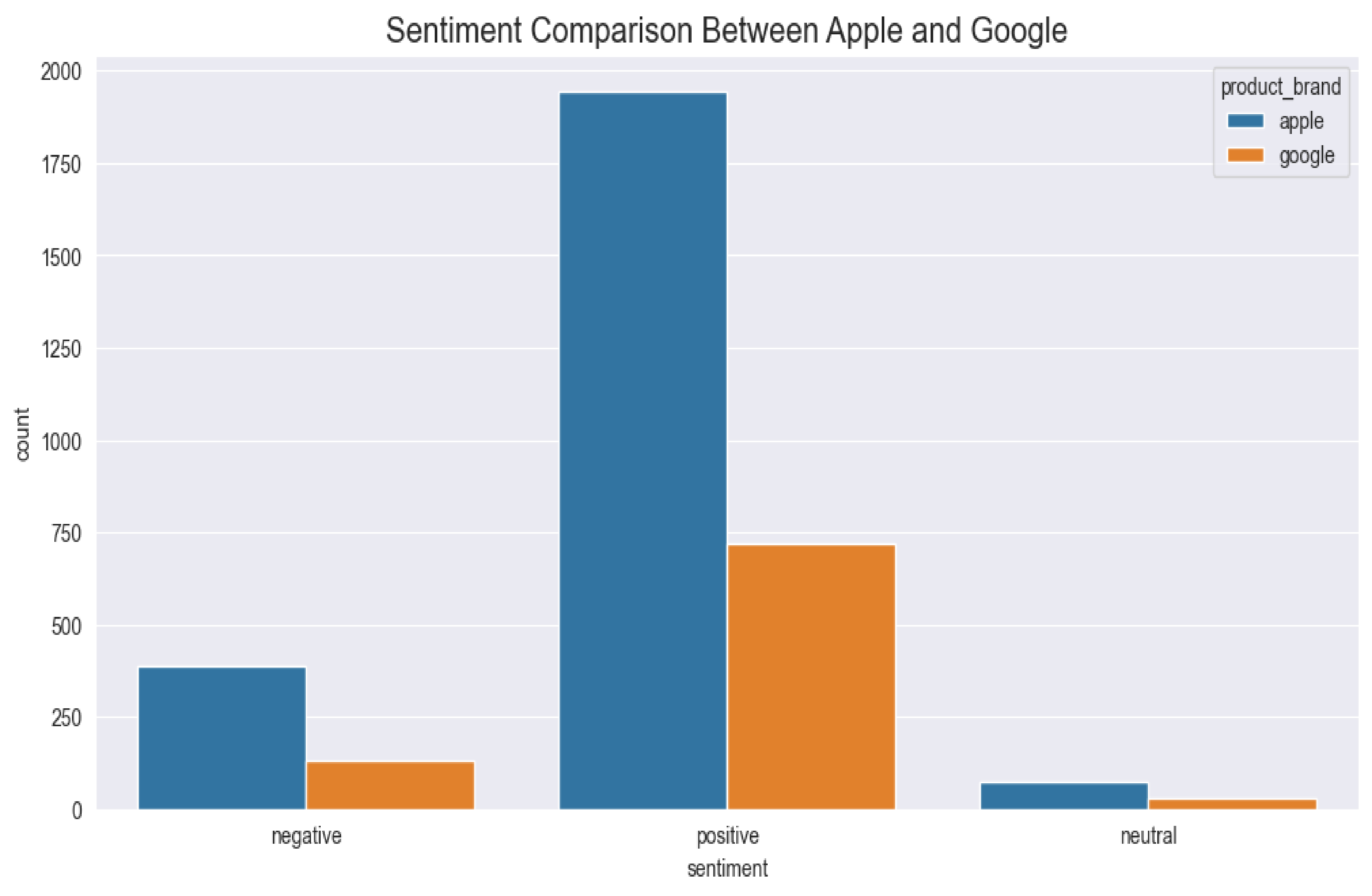
The dataset shows a class imbalance, with more positive and neutral tweets compared to negative ones. This is a common challenge in sentiment analysis tasks.

Modeling Approach

The project initially focused on binary classification (positive vs. negative) and later expanded to a multi-class classification (positive, negative, neutral) to better understand the nuances of sentiment.

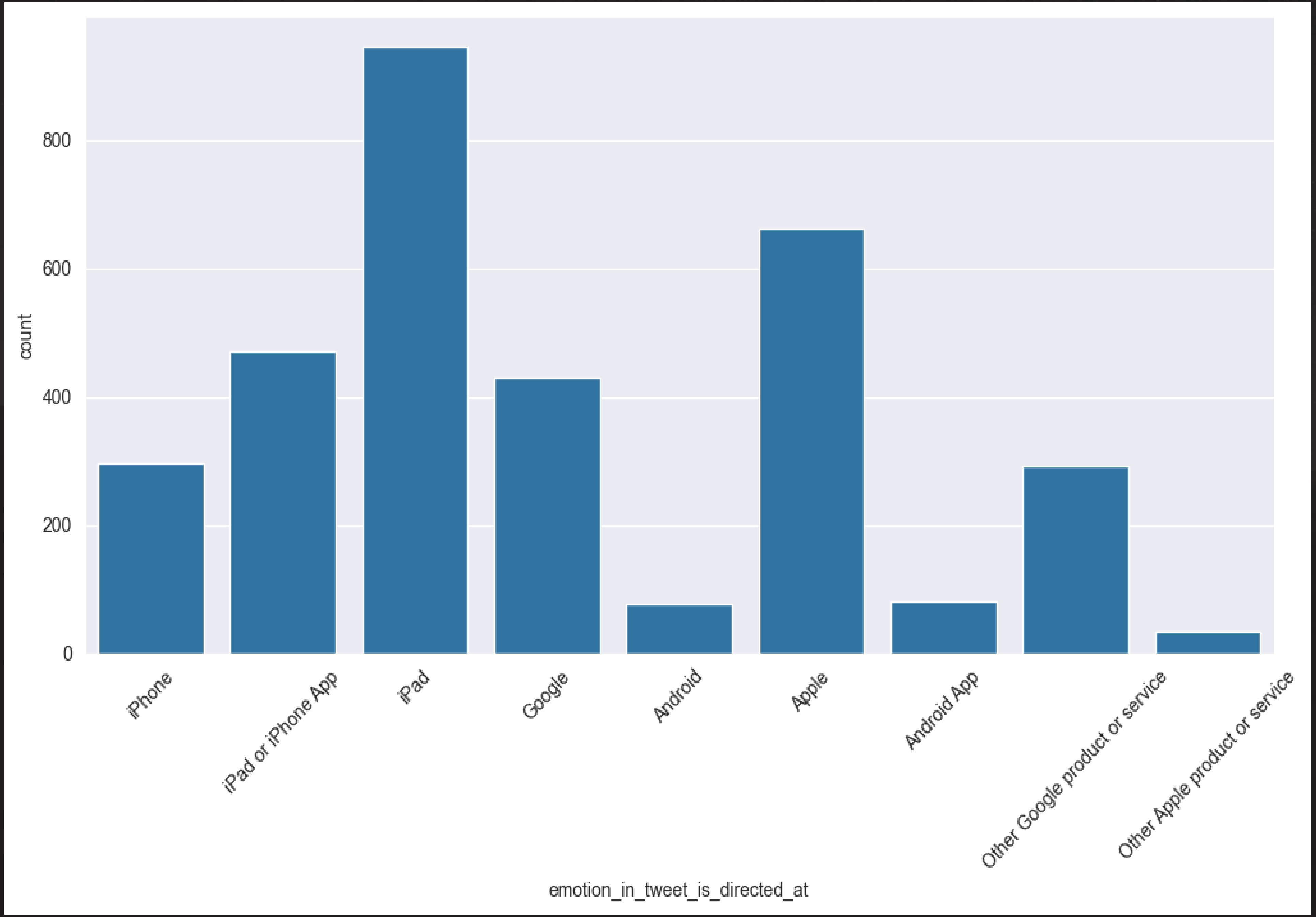
Visualizations

Sentiment Comparison Between Apple and Google Products



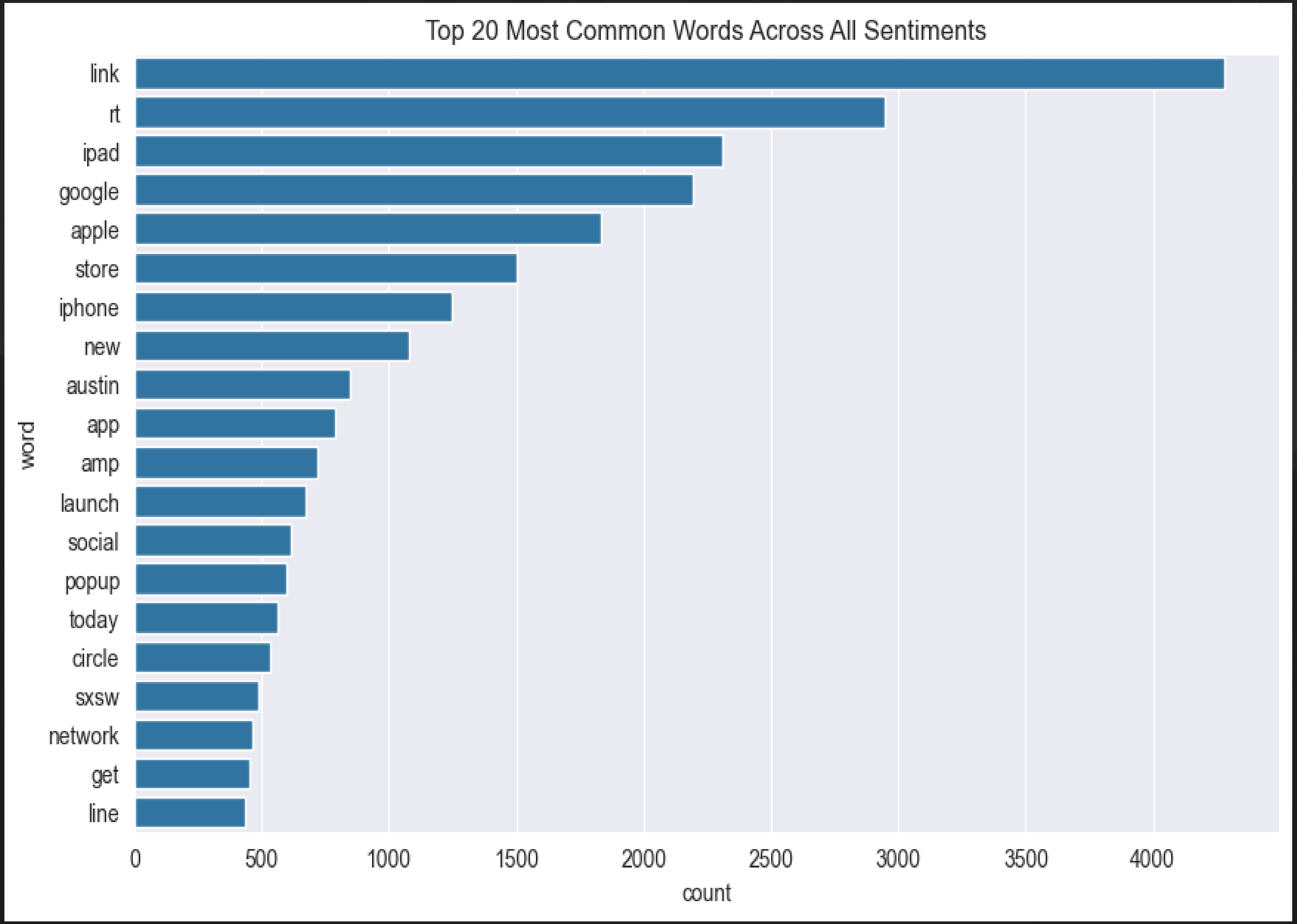
The bar chart compares the sentiment distribution between **Apple** and **Google** products. For **positive sentiment**, Apple has a significantly higher count compared to Google, indicating a strong positive reaction toward Apple products. **Negative sentiment** is more balanced but still higher for Apple than Google. Both brands have very low counts in the **neutral sentiment** category, with Apple showing slightly more mentions than Google. This comparison suggests that Apple products generate more engagement, particularly in positive sentiment, than Google products.

Sentiment Breakdown and Visualization (Continued)



The bar chart reveals that iPad is the most frequently mentioned product in the tweets, followed by other Apple products (iPad, iPhone, and Apple) and Google products. Android-related products receive fewer mentions, highlighting the dominance of Apple products in user-directed sentiments. .

Top 20 Most Common Words Across All Sentiments



- ❑ Here, we combine the preprocessed text from all sentiment categories (positive, negative, and neutral) to analyze the most frequent words across the entire dataset. After calculating word frequencies, we plot the top 20 most common words.
- ❑ From above we notice that common terms like "link", "rt", "ipad", and "google" dominate the conversation, reflecting the general focus of discussions in the tweets

Modeling

Preparing Data for Binary or Multi-class Classification

- We begin by preparing the data for modeling. Depending on the task, we either restrict the dataset to only positive and negative sentiments for binary classification or include neutral sentiments for multi-class classification. We also encode the target sentiment labels into numerical values and prepare the features, including the processed text and product brand.

Vectorization using Tfi-df

- The textual data is transformed into numerical form using TfidfVectorizer, which converts the preprocessed_text into a matrix of TF-IDF features. The categorical variable product_brand is also encoded using OneHotEncoder to incorporate brand information into the model. A ColumnTransformer is used to apply these transformations to the respective features

Pipelines(Binary Classification)

To streamline the process of preprocessing and model training, we define several pipelines for different machine learning algorithms.

These pipelines include:

- ❑ Logistic Regression
- ❑ Random Forest
- ❑ Support Vector Machine (SVM)
- ❑ Gradient Boosting
- ❑ Neural Networks (MLPClassifier)
- ❑ XGBoost

Modeling Techniques

1

Binary Classification

The team evaluated several models for binary classification, including Logistic Regression, Random Forest, SVM, Gradient Boosting, Neural Networks, and XGBoost.

2

Multi-class Classification

For the multi-class classification task, the same set of models were tested to identify positive, negative, and neutral sentiments.

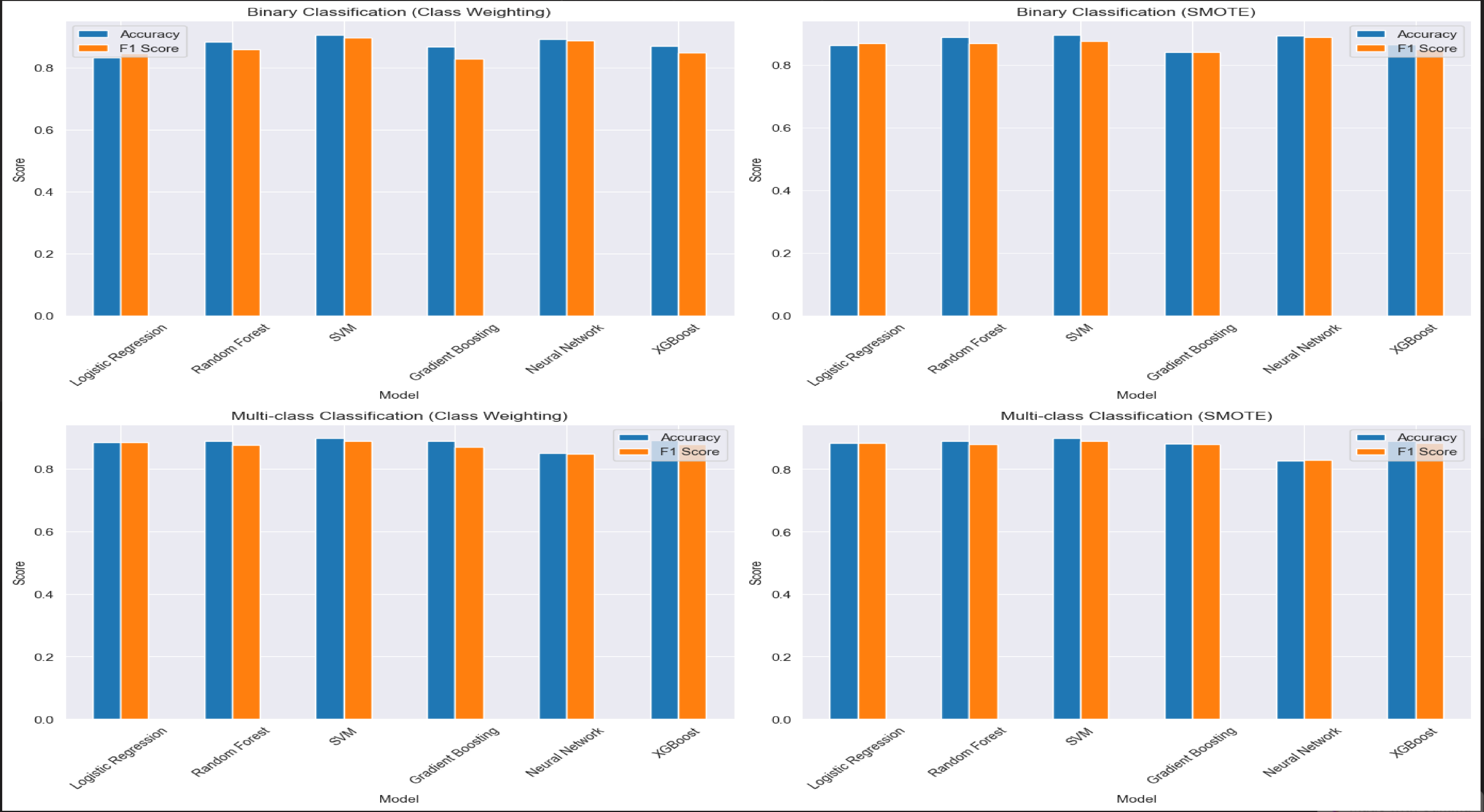
3

Class Imbalance Handling

To address the class imbalance, the team used techniques like class weighting and SMOTE (Synthetic Minority Over-sampling Technique) to improve model performance, especially for the minority class (negative sentiment).

Comparing Model Performance Across Binary and Multi-class Classification

- ❖ **Across both binary and multi-class classifications, SVM is the best-performing model, achieving the highest accuracy and F1 scores. It is closely followed by Gradient Boosting, which also shows strong performance. These two models consistently outperform others in both class weighting and SMOTE settings, making them the most reliable choices for sentiment classification in this dataset.**


Binary Classification (Class Weighting)
Binary Classification (SMOTE)
Multi-class Classification (Class Weighting)
Multi-class Classification (SMOTE)
Accuracy
F1 Score
Score
0.0
0.2
0.4
0.6
0.8
Model
Logistic Regression
Random Forest
SVM
Gradient Boosting
Neural Network
XGBoost

Model Evaluation

1. Binary Classification

In binary classification, SVM with class weighting performed the best, achieving an accuracy of 0.9054 and an F1 score of 0.8973

2. Multi-class Classification

For multi-class classification, SVM with class weighting again showed the best results, with an accuracy of 0.8970 and an F1 score of 0.8903.

3. Class Imbalance Impact

Handling class imbalances significantly improved model performance, especially for the minority class (negative emotions).

4. Comparison of Techniques

Class weighting slightly outperformed SMOTE in most cases, making it the preferred technique for handling class imbalances

Key Observations

Class Imbalance: There seems to be a significant class imbalance in the dataset. Neutral emotions are likely the most common, followed by positive, with negative emotions being rare.

❑ Difficulty with Negative Emotions: All models struggle to identify negative emotions accurately, especially in the binary classification task. This could be due to the class imbalance or the complexity of identifying negative sentiments.

❑ Model Performance: In binary classification, Random Forest and Neural Network perform best overall. In multi-class classification, Logistic Regression, SVM, and XGBoost perform similarly well.

❑ Multi-class vs Binary: The models seem to perform slightly better in the multi-class scenario, possibly because the addition of the neutral class helps to separate positive and negative emotions more effectively.

Recommendations

1

Primary Model Selection

SVM with class weighting is the preferred model for both binary and multi-class sentiment analysis tasks based on the evaluation results.

2

Class Imbalance Handling

Use class weighting as the primary technique for addressing class imbalances, as it slightly outperformed SMOTE in most cases.

3

Further Data Collection

Collect a larger and more balanced dataset to improve the model's performance and generalization, as the current dataset has limitations in size and class

Conclusion

Best Model

SVM with class weighting provided the highest accuracy and F1 scores for both binary and multi- class classifications, making it the best model for this sentiment analysis task.

Class Imbalance Impact

Handling class imbalances significantly improved model performance, particularly for the minority class (negative sentiment), highlighting the importance of addressing this challenge.

Textual Variations and Context

The model faces challenges in handling the informal language and sentiment subtleties commonly seen in social media posts, emphasizing the need for more context-aware models.

Future Improvements

Expand Data Sources

Incorporate more diverse data sources beyond Twitter, such as reviews, forums, and other social media platforms, to enhance the model's ability to handle a wider range of textual variations and contexts.

Contextual Understanding

Explore advanced NLP techniques, like transformer-based models (e.g., BERT, RoBERTa), to better capture the nuances of sentiment and improve the model's understanding of the underlying context.

Real-Time Processing

Develop a scalable solution that can process large volumes of data in real-time, enabling businesses to respond promptly to customer feedback and optimize their marketing strategies based on up-to-date sentiment analysis.

Ethical Considerations

1

Privacy and Data Protection

Ensure that the sentiment analysis model and its deployment adhere to strict data privacy and protection guidelines, respecting the privacy of individuals whose tweets are analyzed.

2

Bias Mitigation

Continuously monitor the model for potential biases, and implement strategies to identify and mitigate any biases that may arise from the dataset or the model's architecture.

3

Responsible AI Practices

Adopt responsible AI principles, such as transparency, accountability, and fairness, to build trust and ensure the ethical use of the sentiment analysis model in business decision-making.

Acknowledgments

CrowdFlower Dataset

The project team would like to acknowledge the CrowdFlower team for providing the valuable dataset used in this sentiment analysis study.

Research Collaborators

The team would also like to thank the research collaborators and subject matter experts who provided valuable insights and guidance throughout the project.

Continuous Learning

The team is committed to ongoing learning and improvement, and welcomes feedback and suggestions from the broader community to enhance the sentiment analysis model and its applications.

Thank You!