Twitter Sentiment Analysis and Classification of tweets on brands

Introduction

In modern digital technology, Twitter has become a popular platform where users openly express their opinions about tech brands and products. Apple and Google, being two of the world's most dominant technology companies, are frequently mentioned in tweets that contain praise, complaints, comparisons or neutral commentary. This project aims to build a sentiment analysis model that can automatically analyze tweets related to Apple and Google products, and classify them as positive, negative or neutral. The insights generated from this classification will help businesses better understand public opinion and improve decision making in areas such as marketing and product development.

Problem Statement

With thousands of daily mentions, Apple and Google receive significant feedback from customers on Twitter. However, analyzing this data manually is both time-consuming and inefficient due to its unstructured nature. Tweets often include informal language, abbreviations, emojis, making them difficult to interpret using traditional rule-based sentiment methods. To solve this, we propose an NLP-driven solution that can automatically and accurately classify sentiments in tweets as positive, negative or neutral. This will empower stakeholders to monitor public perception in real-time, respond proactively to trends and make decisions rooted in customer sentiment.

Objectives

1. Sentiment Analysis: To classify tweets related to technology brands (Apple, Google, or none) into sentiment categories (Positive, Negative, Neutral) and understand public perception.

2. Brand Engagement Analysis: To identify which brands receive more attention and emotional engagement on Twitter, and to compare sentiment distributions across brands.

3. **Binary classification modeling**: To build, tune, and evaluate various machine learning models (Logistic Regression, Decision Trees, Random Forest to classify tweets as either positive or negative

4. **Multiclass classification modeling** - To group the remaining emotion categories as 'neutral' then train models to further classify tweets as neutral

Data Understanding

This project uses a labeled dataset sourced from CrowdFlower via Data.world, containing 9093 tweets primarily focused on Apple and Google products. Each tweet has been manually annotated by human raters who classified each tweet's sentiment and identified the brand or product targeted.

Key columns in the dataset include:

- tweet_text: the content of the Tweet
- emotion_in_tweet_is_directed_at: identifies the targeted product or brand
- is_there_an_emotion_directed_at_a_brand_or_product: indicating the sentiment toward the brand.

Data Analysis

Stacked barchart of Emotion by Brand
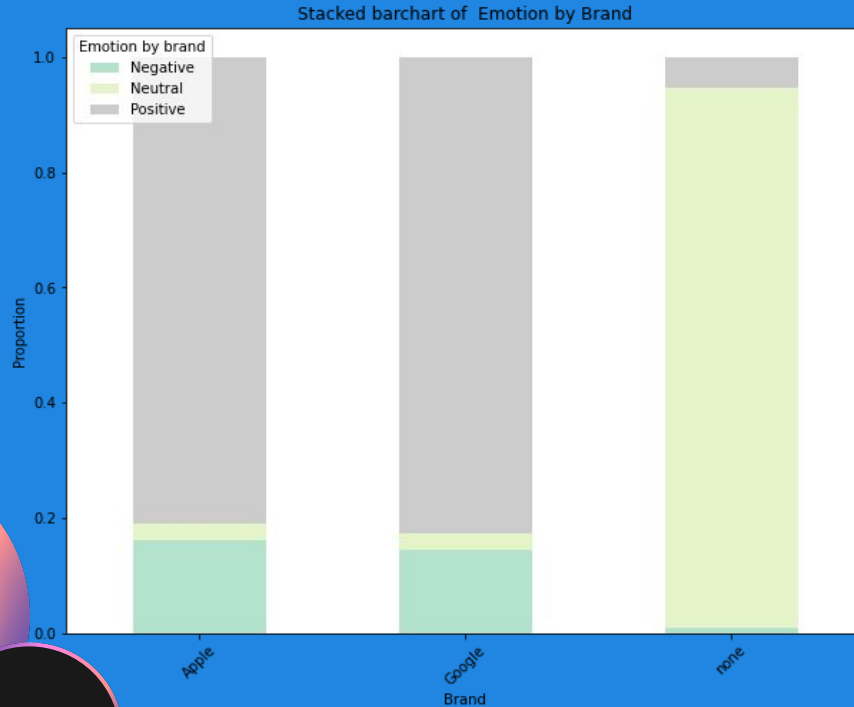
Observations

- Apple and Google Receive Mostly Positive Sentiment
- Apple Has Slightly More Negative Sentiment Than Google
- When no specific brand is mentioned, tweets are mostly neutral, with very little emotion expressed.
- Very Few Positive Tweets in the none category.

Data Analysis
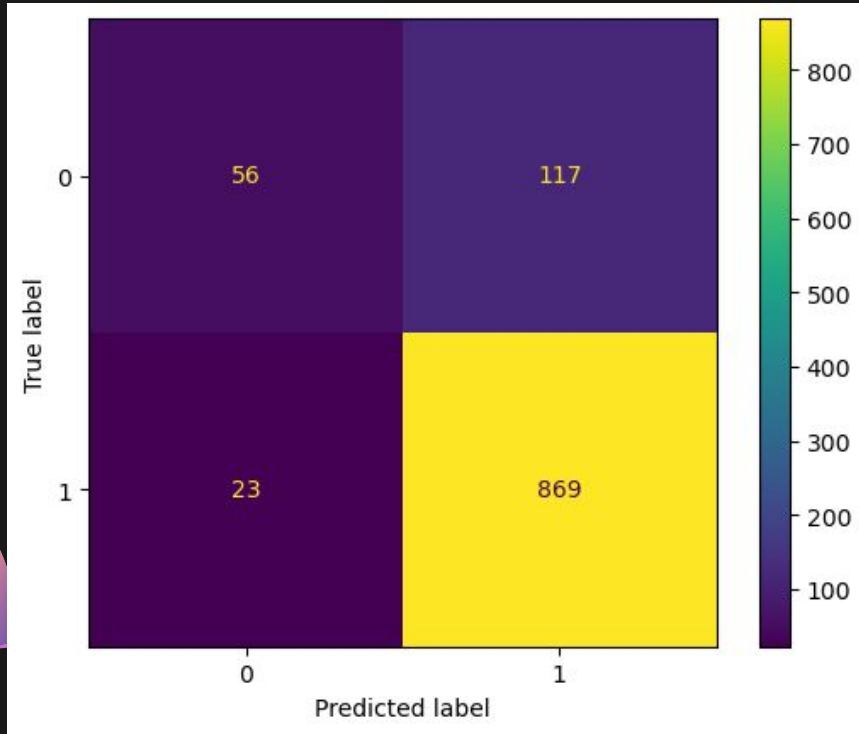
The model performs very well on class 1 (positive/majority class) with high true positives and low false negatives. However, it struggles with class 0 (negative/minority class), misclassifying nearly 68% of them. This indicates class imbalance We'll try fitting a decision tree while accounting for the class imbalance to see if there will be better performance.

Data Analysis
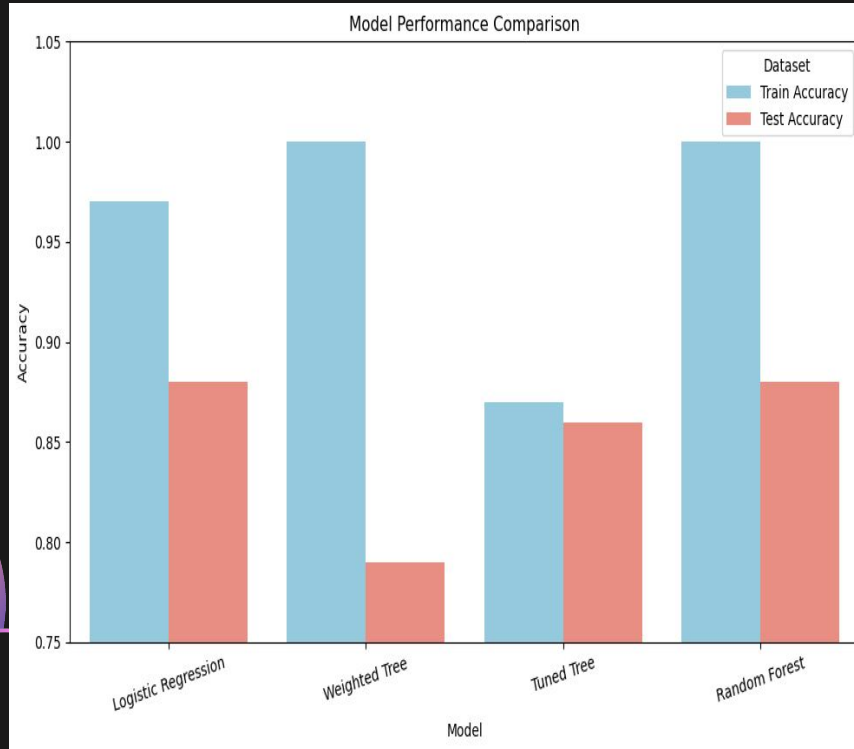
Logistic regression and the tuned tree are our best binary classification models because they do not over fit.

Conclusion

The sentiment analysis project has successfully demonstrated the potential of machine learning models to classify tweets related to Apple and Google products into positive, negative, and neutral sentiments. The binary classification models, particularly Logistic Regression and Decision Trees, provided a solid foundation for understanding sentiment distribution. However, challenges such as class imbalance and overfitting were evident, especially in the more complex models like Random Forest and XGBoost.

The multiclass classification models, including Multinomial Naive Bayes and XGBoost, showed promise in handling multiple sentiment classes but also faced similar challenges. The use of techniques like SMOTE for oversampling did not yield significant improvements, indicating that the dataset's quality and complexity may be limiting factors.

Thanks

Does anyone have any questions?

Contributors:
Julie Maina
Olin Wachira
David Mburu
Denzel Tero
Lyster Kwamboka