TWITTER SENTIMENT ANALYSIS ON APPLE AND GOOGLE PRODUCTS

Prepared by:

- Adrianna Ndubi
- Caleb Kipkech
- Gerald Muhia
- Keith Kimani
- Sarah Joshua

Contents

Twitter Sentiment Analysis on Apple and Google Products

Overview and Business Understanding

Apple is a technological company known for its premium consumer electronics, key products include the iPhone, iPad, Mac computers, Apple Watch, and various services like Apple Music and iCloud. Google, on the other hand, offers a wide range of products including the Android OS, Google Pixel smartphones, Google Nest smart home devices, and services such as Google Drive and YouTube.

Apple focuses on **high-end hardware** integrated with a cohesive software ecosystem, while Google provides a **mix of hardware and software solutions**, emphasizing accessibility and cross-platform integration. Both companies compete in multiple markets, particularly in smartphones **(iPhone vs. Pixel)**, tablets (**iPad vs. Android tablets),** and smart home devices **(Apple HomeKit vs. Google Nest)**. Their rivalry extends to digital services and software ecosystems, with Apple promoting a closed ecosystem and Google advocating for open-source solutions.

As a result of this competition, different stakeholders in the industry undertake sentiment analysis to assist in their respective area of expertise as shown below:

1. **Marketers**: Develop campaigns based on sentiment analysis insights to resonate with target audiences.
2. **Product Teams**: Focus on product development aligned with customer feedback derived from sentiment analysis.
3. **Retailers**: Engage directly with customers and require insights to optimize sales strategies based on sentiment trends.

The proposed solution involves conducting a Twitter sentiment analysis on tweets related to Apple and Google products. This includes building a sentiment classification model to rate tweets as positive, negative, or neutral as well as analyzing consumer sentiment trends over time to identify strengths and weaknesses in product offerings. Insights from the analysis will enable stakeholders to make informed decisions regarding marketing strategies, product development, and customer engagement. Understanding customer sentiment can enhance brand loyalty, improve product offerings, and tailor marketing campaigns.

Problem Statement

At **Group Six Company**, we believe in the power of understanding people. Social media has become a space where consumers freely share their thoughts, opinions, and feelings about the products they use. For a brand like Apple, this feedback is invaluable in staying connected with its audience and delivering what customers truly want.

Apple, a global leader in technology and innovation, has tasked us with analyzing public sentiment about its products using data from Twitter. This project aims to uncover actionable insights by building a robust Natural Language Processing (NLP) model capable of classifying sentiment in tweets as positive, negative, or neutral. Analyzing a dataset of over 9,000 tweets rated by human annotators from CrowdFlower via data. world, we aim to identify key trends and sentiments that can inform Apple's marketing strategies and product development decisions.

Through this analysis, we aim to give Apple a deeper understanding of its customers' perceptions, enabling the company to refine its communication strategies, enhance customer satisfaction, and maintain its position as a market leader in the tech industry.

Objective

Main Objective

- Build a Natural Language Processing (NLP) model that can rate the sentiment of a Tweet based on its content.

Secondary Objectives

- Analyze and Compare Sentiment for Apple vs. Google Products
- Identify Key Drivers of Positive and Negative Sentiment
- Monitor and Track Sentiment Trends Over Time
- Gather insights into customer preferences, opinions, and emerging trends

Metrics of Success

For our analysis, our metrics of success are as follows:

- **Accuracy**: measures the proportion of correctly classified tweets out of the total tweets. It gives a general sense of the model's performance, but it may not always be sufficient in cases of imbalanced datasets.
    - ***Accuracy****= (True Positives + True Negatives) / (Total Samples)*
    - Achieve a classification accuracy of at least **85%** on a balanced test dataset.

- **Recall**: measures the ratio of correctly predicted positive observations to all the actual positives. It is critical when missing positive cases (false negatives) is costly.
    - ***Recall***= *(True Positives) / (True Positives + False Negatives)*
    - Achieve a classification recall of at least **85%** on a balanced test dataset
- **Precision**: measures the ratio of correctly predicted positive observations to the total predicted positives. It is a measure of the model's exactness and is especially important when false positives are more problematic.
    - ***Precision***= *(True Positives + False Positives) / (True Positives)*
    - Achieve a precision of at least 85% on a balanced test dataset.
- **F1 score**: refers to the harmonic mean of precision and recall, balancing the two metrics. It is especially useful when the dataset is imbalanced and when both false positives and false negatives are important.
    - ***F1 Score***= *2 × ((Precision × Recall) / (Precision + Recall))*
    - Target an **F1-score** of 0.85 or higher for each sentiment category to ensure balanced performance.

Data Understanding

The Twitter Dataset used in in this analysis has been sourced from **CrowdFlower** accessible through the link below:

https://data.world/crowdflower/brands-and-product-emotions

It contains roughly 9,000 Tweets which have been evaluated for sentiment by human raters, with the majority focusing on products from Apple and Google. Each tweet has been classified as either positive, negative or neutral. The dataset was collected during the 2013 South by SouthWest (SXSW) conference which is known for showcasing the latest technology. As such, it not only provides a unique setting to effectively capture the real time consumer reactions to brands but also provides an environment where customers can compare products from leading companies with reduced individual biases.

The dataset is made up of 9,093 rows, with each row indicating an independent sentiment review. In addition, it has 3 columns listed as follows:

1. **tweet_text** - Gives a description of the tweet with the corresponding username
2. **emotion_in_tweet_is_directed_at** - Shows the product being referred to by a tweet

3. **is there_an_emotion_directed_at_a_brand_or_product** - Highlights the sentiment i.e. positive, negative or neutral

Data Preparation and Cleaning

The data cleaning process prepares our dataset for analysis by addressing inconsistencies and ensuring that the data is in a usable format. The steps taken include renaming columns, handling duplicates, addressing missing values, merging sentiments, and preparing text data for further exploration. Below, we detail each of these steps and the rationale behind them.

1. Renaming Columns

To make the dataset more manageable, column names were renamed as follows:

- **Tweet**: The column containing the text of the tweets was renamed to tweet_text.
- **Product_Name**: The column emotion_in_tweet_is_directed_at was renamed to reflect the product associated with the tweet.
- **Sentiment_Type**: The column is_there_an_emotion_directed_at_a_brand_or_product was renamed to better describe the type of sentiment associated with each tweet.

This renaming ensures that column names are intuitive and easier to reference in subsequent analysis.

2. Handling Missing Values and Duplicates

a. Checking for Missing Values and Duplicates

- A check for missing values and duplicate rows was conducted to identify issues that might skew the analysis.

b. Handling Duplicates

- Duplicates were removed from the dataset to ensure that repetitive entries do not bias results.

c. Handling Missing Values

- A single missing tweet was removed as its impact on the analysis was deemed negligible.
- For missing values in the **Product_Name** column, a new **Product** column was created. This column assigns:
    - Apple-related products (e.g., iPad, iPhone) to Apple.
    - Google-related products to Google.
    - Unknown products were assigned the label Unknown.

- A function was developed to extract the product name from the **tweet_text** column if the product remained unknown. This extracted name was stored in the new **Brand** column.

3. Merging Sentiments

- To simplify sentiment analysis, the **Sentiment_Type** column's labels were consolidated into three main categories:
    - Positive
    - Negative
    - Neutral: This includes tweets labelled "I can't tell," which were merged into the neutral category due to their low occurrence.

4. Text Data Preparation and Analysis

Text data from the Tweet column was prepared using several Natural Language Processing (NLP) techniques:

a. Initial Preprocessing

1. **Standardizing Text:**
    - All tweets were converted to lowercase for uniformity.
    - A list of tweets was created for tokenization.
2. **Tokenization:**
    - A tokenizer specifically designed for Twitter, TweetTokenizer was used to dissect tweets into tokens.
3. **Hashtag and Accent Removal:**
    - Hashtags and accents were removed from tokens to focus on meaningful words.

b. Removing Noise

1. **Punctuation Removal:**
    - Punctuation marks were eliminated from the tokens to reduce noise in the text.
2. **Stop word Removal as well as removal of unwanted words:**
    - Common stop words were removed from the tokenized text.
    - Additional irrelevant words (e.g., "sxsw," "link," and "RT") were excluded. The number "2" was also removed as it primarily referred to the iPad 2 and was generalized to "iPad."

    - Special characters (e.g., "\x89") and single-character tokens were removed using regular expressions.

c. Lemmatization

- A lemmatizer with part-of-speech (POS) tagging was used to convert words to their base forms. This step ensured that variations of a word (e.g., "running," "runs") were treated as the same term.

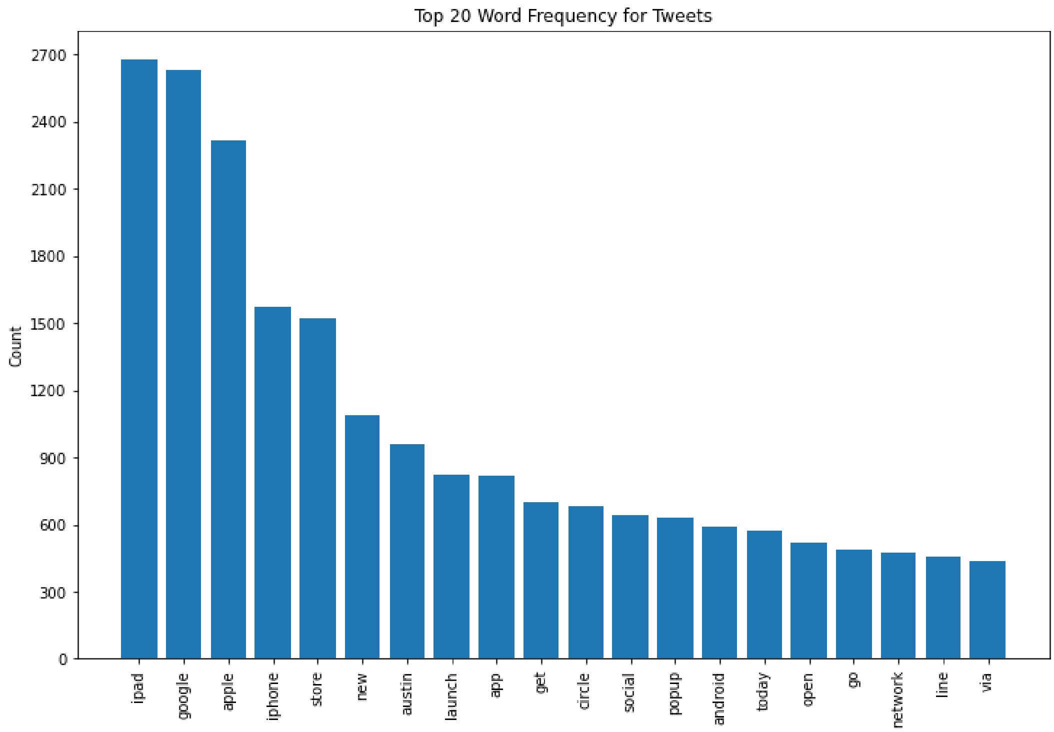
d. Token Analysis

- Frequency distributions of the cleaned tokens were generated to identify the most frequently occurring words. After preprocessing, "iPad" emerged as the most frequent word, followed by "Google."

5. Automating the Process

- A function was created to streamline the entire text preprocessing workflow for the Tweet column. The steps included:
    - Tokenization
    - Removal of hashtags and accents
    - Elimination of punctuation
    - Mapping of NLTK POS tags to WordNet POS tags
    - Lemmatization with POS tagging
    - Removal of stop words, irrelevant terms, and special characters

The resulting top 20 words are as shown in the figure below:

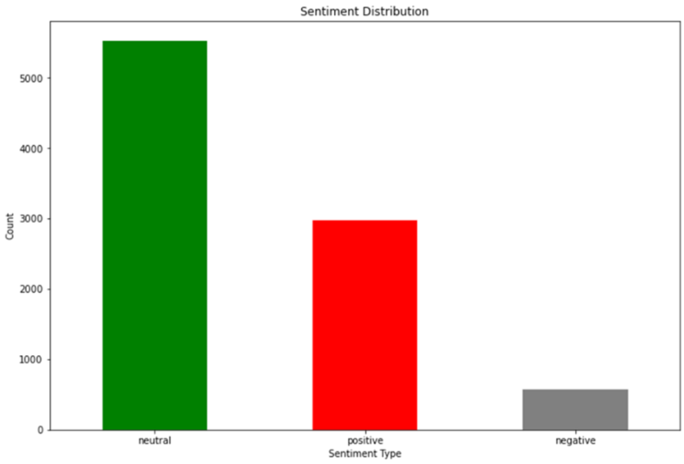


This data cleaning process ensured that the dataset is well-prepared for sentiment analysis and other NLP-based tasks. By consolidating sentiment categories, addressing missing values, and thoroughly cleaning the text data, we have created a dataset that can be used for insightful analysis and modelling.

Exploratory Data Analysis

Sentiment Distribution

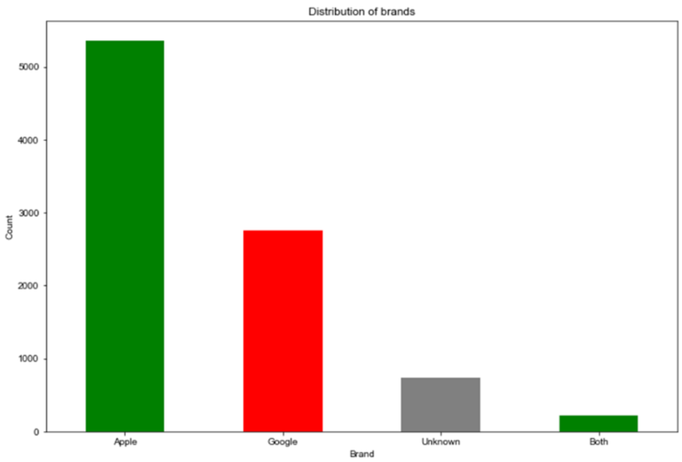
The sentiment distribution graph below reveals the relative prevalence of sentiment types across the dataset.



Neutral sentiments dominate, suggesting that most tweets neither express strong positivity nor negativity. Positive sentiments come in second, indicating a significant number of favourable opinions. Negative sentiments are the least common, implying that fewer users expressed dissatisfaction or unfavourable opinions.

Distribution of brands

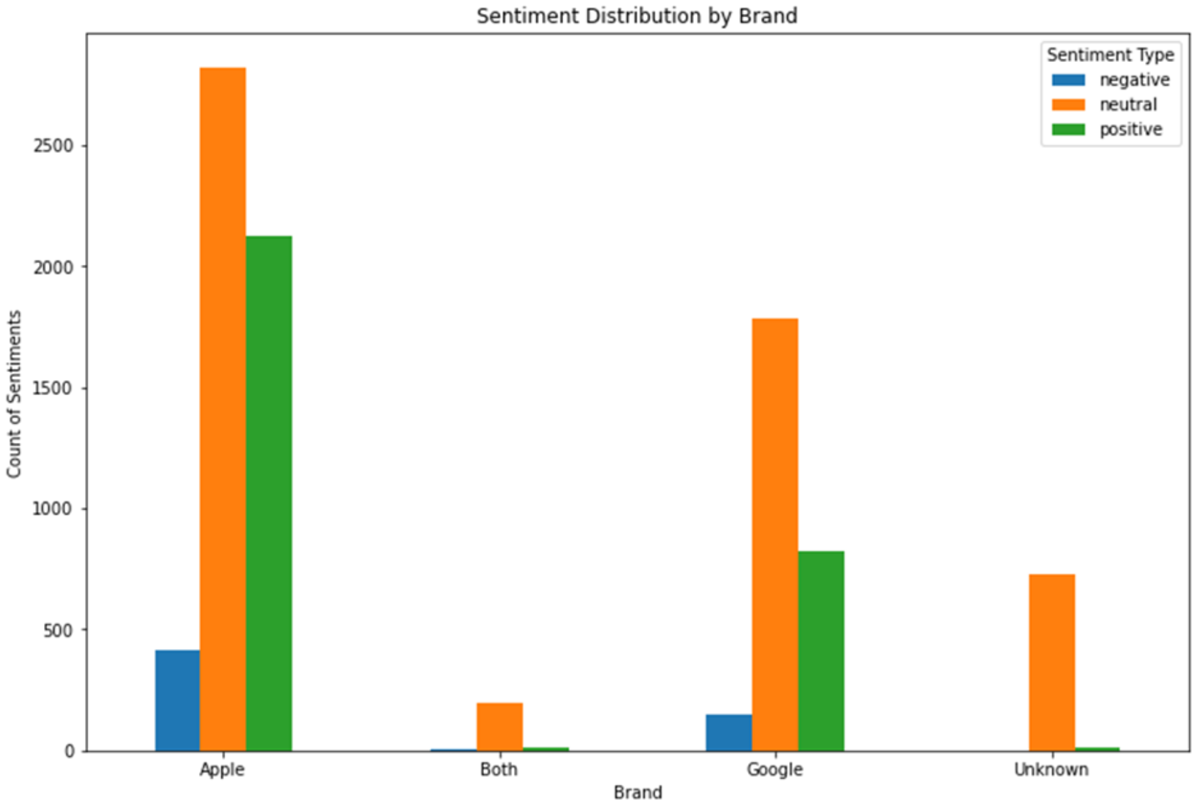
The brand distribution on the following page illustrates the number of tweets mentioning specific brands.



Apple received the highest attention, with over 5,000 tweets, highlighting its prominence during the analyzed period. Google followed with approximately 3,000 tweets. Interestingly, tweets that did not explicitly mention any brand outnumbered those mentioning both brands. The strong focus on Apple suggests either its active presence during the event or a stronger product appeal. This may reflect better marketing strategies, product launches, or consumer interest in Apple's offerings compared to Google.

Sentiment Distribution by Brand

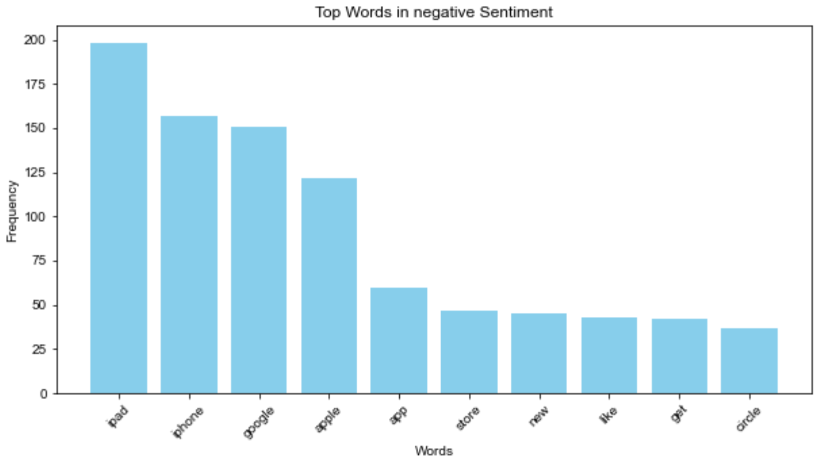
The graph below provides insight into how sentiment varies across brands

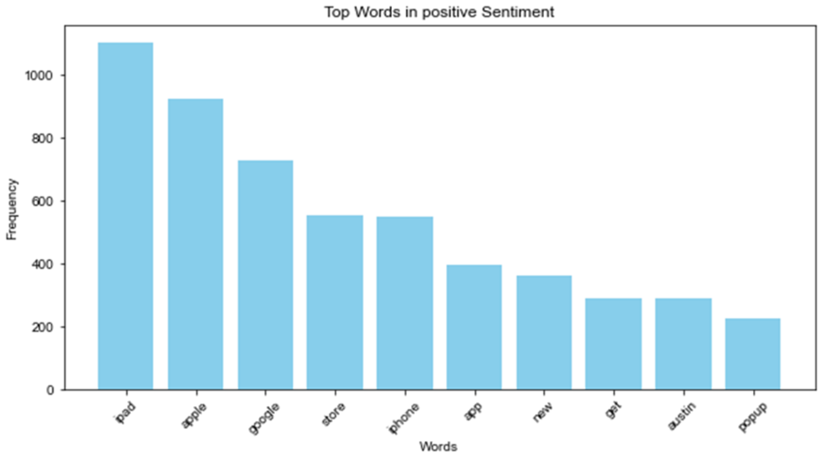


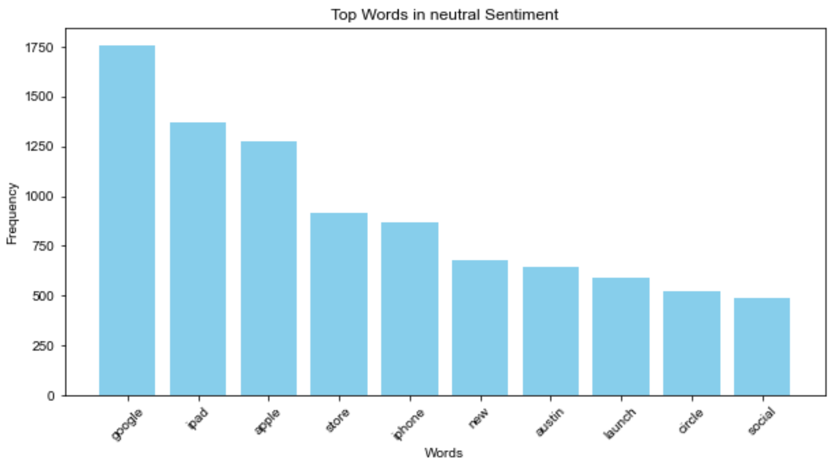
For both Apple and Google, neutral sentiments are predominant, indicating that most users did not express strong opinions. Positive sentiments are more frequent than negative ones for both brands, with Apple showing a slightly better positive sentiment proportion compared to Google. In the Unknown category, neutral sentiments are disproportionately higher, suggesting that tweets not directed at specific brands were more likely to be impartial or lacked emotional context. Apple's edge in positive sentiments suggests a more favourable consumer perception compared to Google.

Top Words per Sentiment

The analysis of top words by sentiment highlights common themes in user opinions as shown by the graphs below:







For positive and negative sentiments, "iPad" emerges as the most frequently mentioned word, signifying its central role in both favourable and unfavourable discussions. For neutral sentiments, "google" appears most often, reflecting its strong association with impartial or general tweets. This pattern may indicate that Apple's products, particularly the iPad, evoked stronger emotional reactions (positive or negative), whereas Google was discussed in a more neutral tone.

Modelling:

Following the exploratory data analysis, next steps is to preprocess in readiness for modelling following the steps listed below:

Preprocessing:

1. *Feature Engineering*
    - Converted lemmatized tokens to strings and stored them in a new column "Processed_Tweets".
    - Created a new dataframe "binary_df" for binary classification (Positive vs. Negative).
    - Created a new column "y_binary" in the “binary_df” with 1 for positive and 0 for negative sentiments.
2. *Check for class Imbalance:*
    - The target variable “y_binary” has a class imbalance (83% positive, 17% negative).
    - Baseline model accuracy would be roughly 83% by always predicting the majority class.
3. *Train-Test Split:*
    - Split data into training and testing sets. Model will be trained on the training data set and tested on the test data set.

Modelling was done in two phases:

i. Binary Classification
ii. Multiclass Classification

Binary Classification:

i. Baseline Model - Multinomial Naïve Bayes:

- Used Multinomial Naive Bayes due to its efficiency and suitability for text data.
- Achieved baseline accuracy of 85%, slightly better than the majority class baseline.
- However low precision of 72% was observed indicating misclassification of negative sentiments.

ii. *Hyperparameter Tuning:*

- Performed Grid Search on model parameters i.e. TF-IDF vectorizer and Multinomial Naive Bayes parameters to seek out the best parameters that would improve the predictive power of the model.

- Improved precision and recall to roughly 88%, resulting in F1-score of 86%.

iii. *Oversampling (Synthetic Minority Oversampling Technique -SMOTE):*

- Given the class imbalance observed in the data, oversampling using SMOTE was done on the data to handle it and likely improve model performance.

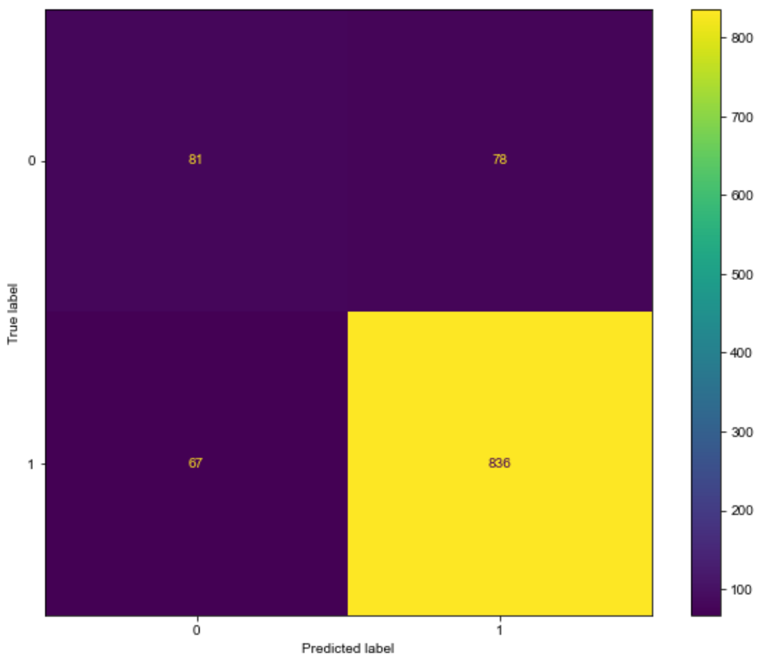
- Oversampling with SMOTE led to a decreased performance, particularly in recall.

iv. *Neural Network*

- Implemented a neural network on the binary classification problem with increased hidden layers, epochs, and smaller batch sizes. Aim was to check if we can improve on the predictability of our outcomes using a different model.

- Achieved slightly lower performance than the best Multinomial Naive Bayes model.

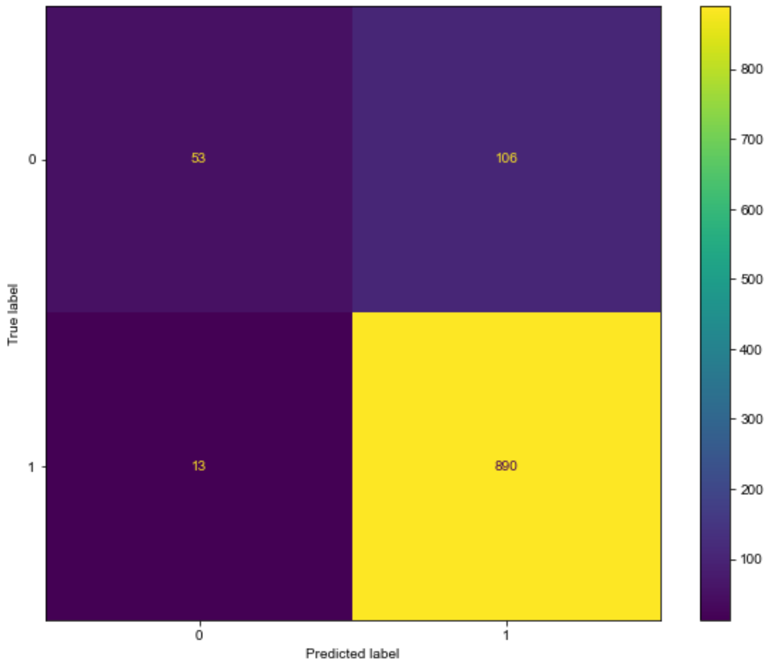
- Confusion matrix showed improved negative class classification but decreased positive class performance.



- . It was noted that the number of correctly classified negatives has increased, indicating better performance for the negative class. There are also fewer false positives in the current matrix, indicating better precision. However, the number of false negatives has increased, meaning the model is now missing more actual positive samples.

v. *Best Model:*

- Tuned Multinomial Naive Bayes model without oversampling achieved the best results.
- Confusion matrix analysis revealed good performance for the positive class but some misclassification of negative instances. The model performs well at correctly classifying positive instances (890). There is a moderate number of samples misclassified as positive (106), suggesting room for improvement in precision. The model rarely misses positive samples (13), which is good for recall. Fewer instances (53) are correctly classified as negative, indicating the negative class might be under-represented or harder to classify.



Multi-class Classification:

i. *Preprocessing*

- To perform this multi classification modelling:
  - Label encoded target variable for three classes (Positive, Negative, Neutral).

- Split the dataset into train and test splits to train and test performance of our models respectively.

ii. Baseline Model:

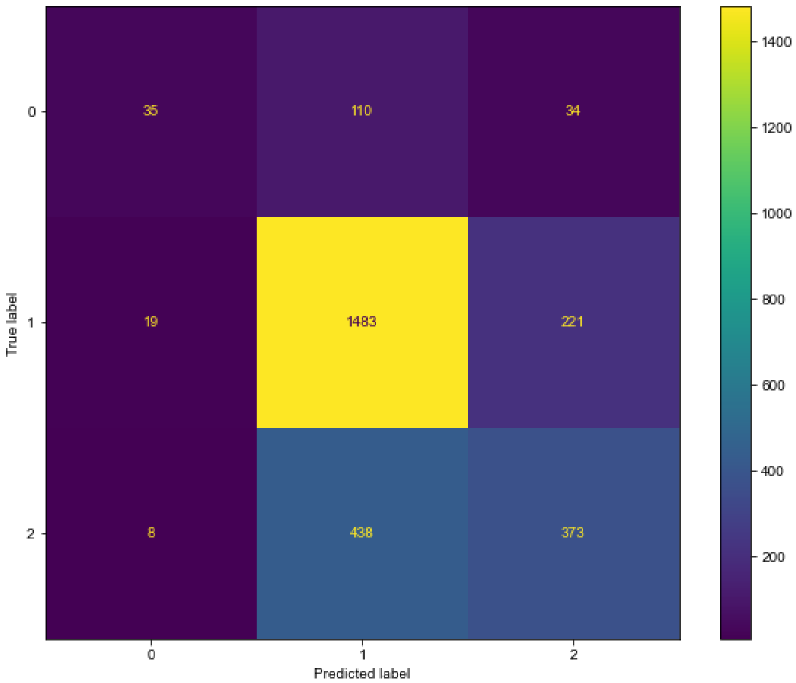
- Multinomial Naïve Bayes was used as the baseline model like the binary classification
- Achieved accuracy of 64% with low precision and recall. This means the model is misclassifying the sentiments i.e. positive classified as either neutral or negative or vice versa. As such, the parameters in the model need to be tuned to achieve a better result

iii. Hyperparameter Tuning:

- Tuning some of the hyper parameters in the model improved the Multinomial Naïve Bayes model performance. Accuracy, precision and recall increased to roughly 68%. This also led to an increased F1 score of 0.67. Model is now performing much better.

iv. XGBoost:

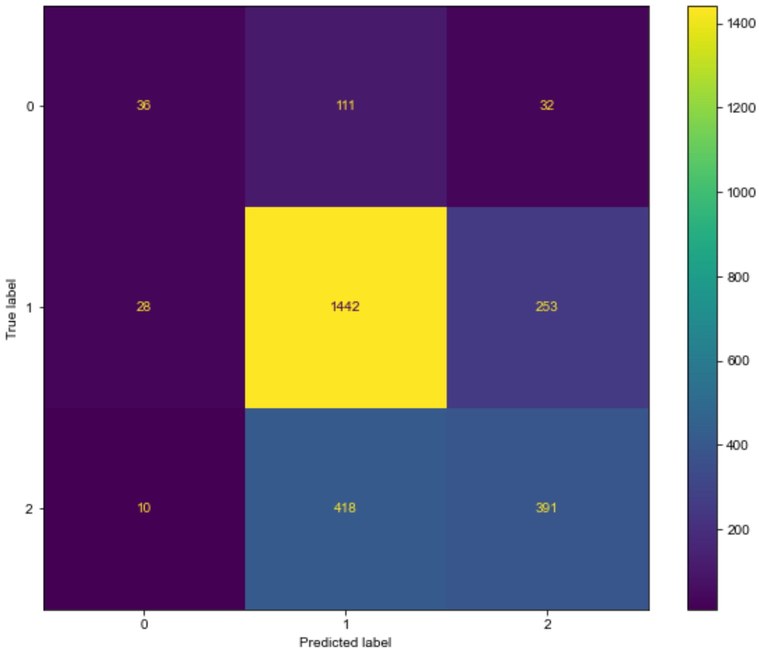
- Employed an additional model XG Boost to improve the predictions of our data as it's suitable for a small dataset like the one used in this analysis.
- Outperformed Multinomial Naïve Bayes with accuracy around 70% and improved precision and recall. Please see the confusion matrix below:



- As per the confusion matrix above, the model correctly predicted 35 instances of Class 0 (Negative), 1483 instances of Class 1 (Neutral) and 373 instances of Class 2 (Positive). Class 1 (Neutral) has significantly more samples compared to Classes 0 (Negative) and 2 (Positive), leading to dominant predictions for Neutral class. Misclassification is heavily skewed against the minority classes (positive and negative). Hence a high recall for Neutral class. The model performs well in identifying Neutral class (1483), likely due to its larger representation in the dataset but has challenges with minority Classes. Positive and Negative classes have higher rates of misclassification, with many instances being confused with the Neutral class. This suggests the model struggles with minority classes and may benefit from strategies to handle class imbalance.

v. *Oversampling (SMOTE):*

- Oversampling slightly improved precision but decreased accuracy and recall.
- Increased misclassification of the Neutral class as shown in the confusion matrix below:



- Class 0 (Negative) and Class 2 (Positive) both show slight improvements in true positive rates. Misclassification of positive instances as neutral has decreased by 20 cases. Class 1 (Neutral) sees a noticeable drop in correct classifications (1442 vs. 1483) and an increase in misclassifications into other classes. Misclassifications of Class 1 (Neutral) and Class 2 (Positive) increased significantly (+32 cases). As a result, the current model slightly balances performance between Classes 0 (Negative) and 2 (Positive) while sacrificing performance for Class 1 (Neutral), indicating a shift in the focus of the model.

Model Evaluation:

Here are the summary findings of both the binary and multi class modelling:

- **Binary Classification:**
  - Tuned Multinomial Naïve Bayes without oversampling was the best model.
- **Multi-class Classification:**
  - XGBoost without oversampling achieved the best results.

Key observations are listed as follows:

- Class imbalance significantly impacted model performance.
- Oversampling did not consistently improve results in both binary and multi-class scenarios.
- Multinomial Naïve Bayes performed well for binary classification.
- XGBoost outperformed Multinomial Naïve Bayes for multi-class classification.
- Further investigation and potential data augmentation techniques are needed to improve performance on the minority classes.

Recommendation

- **Focus on improving performance for negative sentiment:** Both binary and multi-class classification models struggled with accurately identifying negative sentiment. This could be due to the class imbalance in the dataset, where negative tweets are outnumbered by positive and neutral tweets. Techniques like oversampling or exploring different classification algorithms specifically designed for imbalanced datasets might be beneficial.
- **Investigate the impact of SXSW on sentiment:** The data was collected during the SXSW conference, which might not represent typical user sentiment. Consider collecting data from a broader range of sources to get a more comprehensive picture of user opinions.
- **Target specific products for analysis:** This analysis focused on sentiment for Apple and Google products in general. Conducting a more granular analysis on specific products (e.g., iPhone vs. Android) could reveal more targeted insights.
- **Actionable insights for marketing and product development:** The report suggests that Apple received more attention and positive sentiment compared to

Google. It would be beneficial for Apple to understand what drives this positive sentiment.

- **Further explore top words by sentiment:** The analysis of top words by sentiment provides a starting point, but a deeper dive could be beneficial. This could help identify underlying themes and topics associated with positive, negative, and neutral sentiment.

By implementing these recommendations, Apple can gain more valuable insights from their social media data and use them to improve their marketing strategies, product development, and overall customer experience.

Conclusion

This analysis aimed to build a robust NLP model to classify sentiment in Twitter tweets related to Apple and Google products. Through a comprehensive approach that included data cleaning, exploratory data analysis, and model building (including Multinomial Naive Bayes, Neural Networks, and XGBoost), valuable insights were gained into consumer sentiment.

Key findings include the dominance of neutral sentiment across both brands, a higher proportion of positive sentiment for Apple, and the challenge of accurately classifying negative and positive sentiments, particularly in the multi-class classification.

While the Tuned Multinomial Naive Bayes model demonstrated strong performance in binary classification, XGBoost proved more effective for multi-class classification. However, both models faced challenges in accurately classifying minority sentiment classes.

This analysis provides a foundation for understanding consumer sentiment towards Apple and Google products. Further research, including exploring advanced techniques like deep learning models, incorporating external data sources, and addressing class imbalance more effectively, can further enhance the accuracy and insights derived from this analysis. The findings of this study can be valuable for Apple in refining its marketing strategies, improving product development, and enhancing customer satisfaction in the competitive tech landscape.