NLP-tweet-sentiment-project

Collaborators and Author of these Report

Gerald Mwangi

Ian Kiptoo

Lynn Komen

1: Business understanding

1.1: Business overview

Brand and Product Emotions refer to the feelings and attitudes that consumers express towards specific brands and their products. These emotions, which can be positive, negative, or neutral, are influenced by factors such as personal experiences, marketing efforts, product quality, and social trends. . We can offer useful insights for enhancing customer satisfaction and brand reputation by categorizing the emotions conveyed in tweets and identifying the targeted brands or items. Social media sentiment has a big influence on how consumers perceive brands and behave. Negative attitudes can cause reputational harm and consumer attrition, while positive sentiments can increase brand loyalty and draw in new clients. As a result, it is critical for Cadel Electronics Company marketing department to track and evaluate public opinion in order to proactively fix problems and capitalize on favorable comments for advertising.

1.2: Stakeholder

Codel Electronics brand Manager who is responsible for overseeing the brand's image and reputation, ensuring it aligns with the company's values and market goals

1.3: Problem Statement

Codel Electronics Company lacks comprehensive insights into consumer emotions and public opinion towards its products and brands. This hinders their ability to identify areas for improvement, enhance customer satisfaction, and develop effective market strategies. A systematic approach to analyzing Twitter sentiment is needed to gain actionable insights and stay competitive.

1.5: Proposed Solution

Analyzing Twitter sentiment to understand consumer emotions towards Apple and Google products, identify areas for improvement, and enhance market strategies to comprehend public opinion about products and brands offered by Codel Electronics Company.

1.6: Objectives:

1.6.1: Main Objective

To develop predictive models to classify the sentiment (positive, negative, or neutral) expressed in tweets about Apple and google brand and products.

1.6.2: Specific Objectives

To leverage twitter data to manage and comprehend public opinion about google and apple products brands offered by Codel Electronics

To identify the best performing product, in terms of positive and negative

To understanding sentiment (emotion) distribution in regard to apple and google brand

To providing actionable insights from analyzing the twitter sentiments

1.7: Success Criteria

Goal: Creating a multiclass machine learning mode with an accuracy of more than 75% and a f1_score of more than 70%

1.8: Constraints

High computational needs required when tuning and optimizing our models

Quality of data - Target variable class imbalance

2: Data Understanding

The dataset comes from Crowd Flower via data. World Source, It contain 3 columns with over 9000 tweets of people sentiments about google and apple products on twitter which were classified as either positive, negative, no emotions or .I can't tell

2.1: Features (columns)

Tweet text (Categorical): The text of the tweet. This feature contains the actual tweet content posted by users, it's also our predictor variable

Emotion_in_tweet_is_directed_at (Categorical): The brand or product that the emotion in the tweet is directed at. This feature identifies which brand or product is the target of the emotion expressed in the tweet.

Is_there_an_emotion_directed_at_a_brand_or_product (Categorical): Indicates whether there is an emotion directed at a brand or product. The possible values include "Positive emotion", "Negative emotion", and "No emotion toward brand or product", it's also our target variable

2.2: Data Exploration

2.2.1: First 5 Rows:

The dataset comprises tweets about multiple brands and products, along with emotions directed towards them.

2.2.2: Basic Data Information:

The dataset has 3 columns:

. Tweet text: Contains the text of the tweet.

. Emotion_in_tweet_is_directed_at: Indicates the brand or product the emotion is directed at.

. Is_there_an_emotion_directed_at_a_brand_or_product: Specifies whether there is a positive, negative, or no emotion directed at the brand or product.

The dataset has a total of 9093 entries.

. The tweet text column has 9092 non-null values.

. The emotion_in_tweet_is_directed_at column has 3291 non-null values.

. The is_there_an_emotion_directed_at_a_brand_or_product column has 9093 non-null values.

2.2.3: Data Shape:

The dataset consists of 9093 rows and 3 columns, indicating there are 9093 observations and 3 features.

2.2.4: Data Types:

All three columns (tweet text, emotion_in_tweet_is_directed_at, and is_there_an_emotion_directed_at_a_brand_or_product) are of object type, indicating they contain categorical or text data.

2.2.5: Summary Statistics:

The summary statistics provide insights into the distribution of data in the dataset.

Tweet text:

Count: 9092 (total observations)

Unique: 9065 (unique tweets)

Top: "RT @mention Marissa Mayer: Google Will Connect You with the Future!" (Most frequent tweet)

Frequency: 5 (occurrences of the most frequent tweet)

emotion_in_tweet_is_directed_at:

Count: 3291 (non-null observations)

Unique: 9 (unique brands/products)

Top: "iPad" (most frequent brand/product)

Frequency: 946 (occurrences of the most frequent brand/product)

is_there_an_emotion_directed_at_a_brand_or_product:

Count: 9093 (total observations)

Unique: 4 (unique sentiment categories)

Top: "No emotion toward brand or product" (most frequent sentiment)

Frequency: 5389 (occurrences of the most frequent sentiment)

These statistics are useful for understanding the data distribution and identifying any potential inconsistencies or areas for further investigation. The dataset provides a solid foundation for analyzing public sentiment towards brands and products based on Twitter data.

3: Data Preparation

3.1: Data Cleaning

Steps to be followed

3.1.1: Completeness

We checked for completeness of our data by checking for missing values in our data, we found one null value in the tweet_text column and 5370 in the emotion_in_tweet_is_directed_at column after careful investigation we noticed it was an empty tweet which we decided to drop since it won't affect our analysis but for the emotion in tweet is directed at column we left them since we won't be using it for modelling or analysis

3.1.2: Consistency

Consistency in our data is very important hence we had to check for duplicates values in our data and we found 22 records and after investigating further in each column we found the duplicates were as result of repletion of tweets hence we decided to drop them

3.1.3: Uniformity

Viewing our columns name we decided to change them into more simple names for easy of readability and interpretation i.e.; tweet text, emotion in tweet is directed at, is there an emotion directed at a brand or product into tweets, product and emotion respectively

Second, we converted all no emotions towards brand values from emotion column to neutral emotion for easy of interpretation

Third, we did text cleaning by removing punctuations, white spaces, hashtags, symbols and numbers as well as converting uppercase letters to lowercase to prepare our text date for preprocessing

3.1.4: Tokenizing

We performed tokenization on our tweets these is a process of breaking down a pieces of text, like a sentences into individual words or tokens these tokens helps computers understand our text data and process by splitting it into manageable units

3.1.5: Removing of stop words

We removed stop words in build in nltk library as well these additional words to our stop words ; 'sxsw', 'mention', 'link', 'rt', 'app', 'android', 'sxswi', 'party', 'mobile', 'apps', 'downtown', 'maps', 'check', 'mayer', 'marissa', 'googles', 'us', 'pop', 'news', 'win', 'first', 'launch', 'panel', 'shop', 'booth',

'apples', 'itunes', 'ipads', 'blackberry', 'temp', 'designing', 'tv', 'fb', 'quotgoogle', 'uberguide', 'ubersocial', 'gsdm', 'interactive', 'flipboard', 'tapworthy', 'sampler','navigation', 'quotthe', 'qagb', 'foursquare', 'wifi', 'hootsuite', 'checkins' these enables us to shorten our texts for easy processing and modeling as well as removing irrelevant words that might affect our modeling

3.1.6: Normalization

We normalized to reduce the words to their root form hence we choose lemmatization so as to retain most of the information and meaning of the words

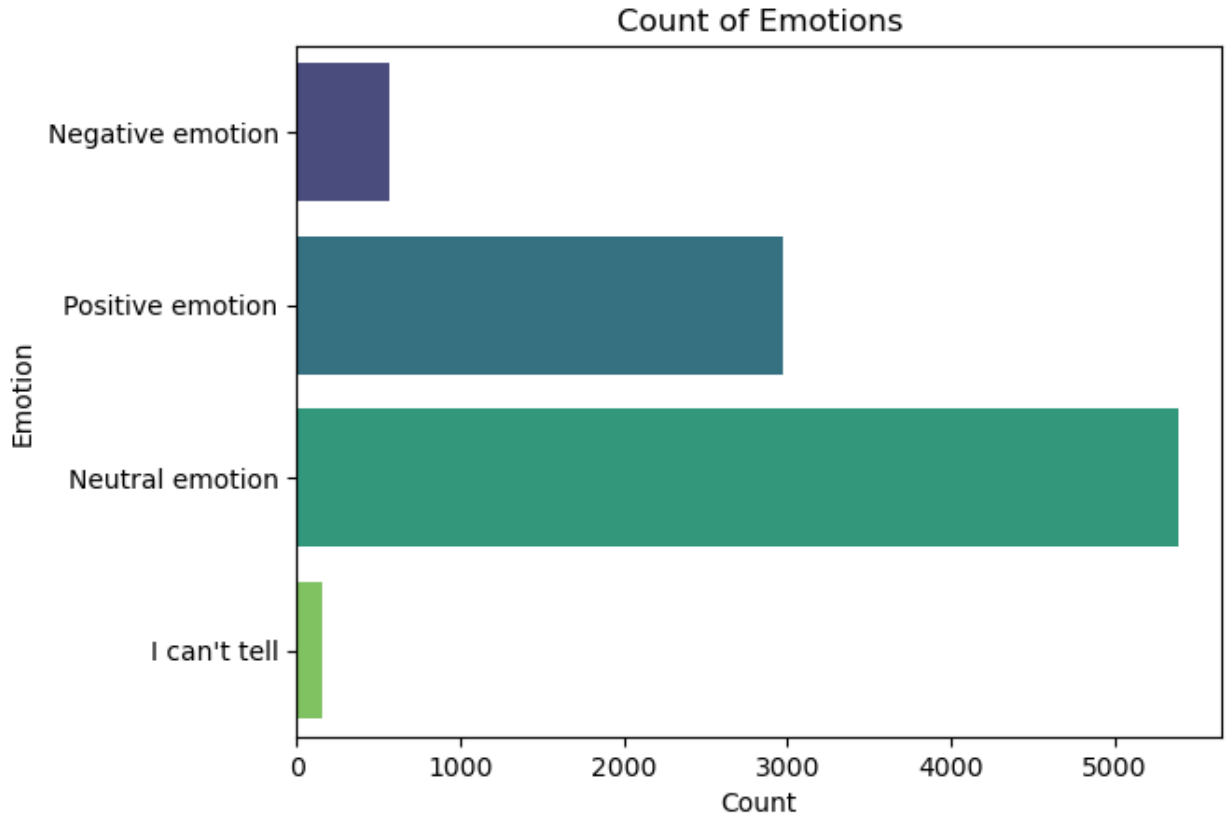
3.2: Feature engineering

We created a new feature company_name which enables us to view the different products mentioned in the tweets either there from Apple or Google which could be usefully in analysis of the better loved brand

4: EDA and Data visualization

4.1: Univalent analysis

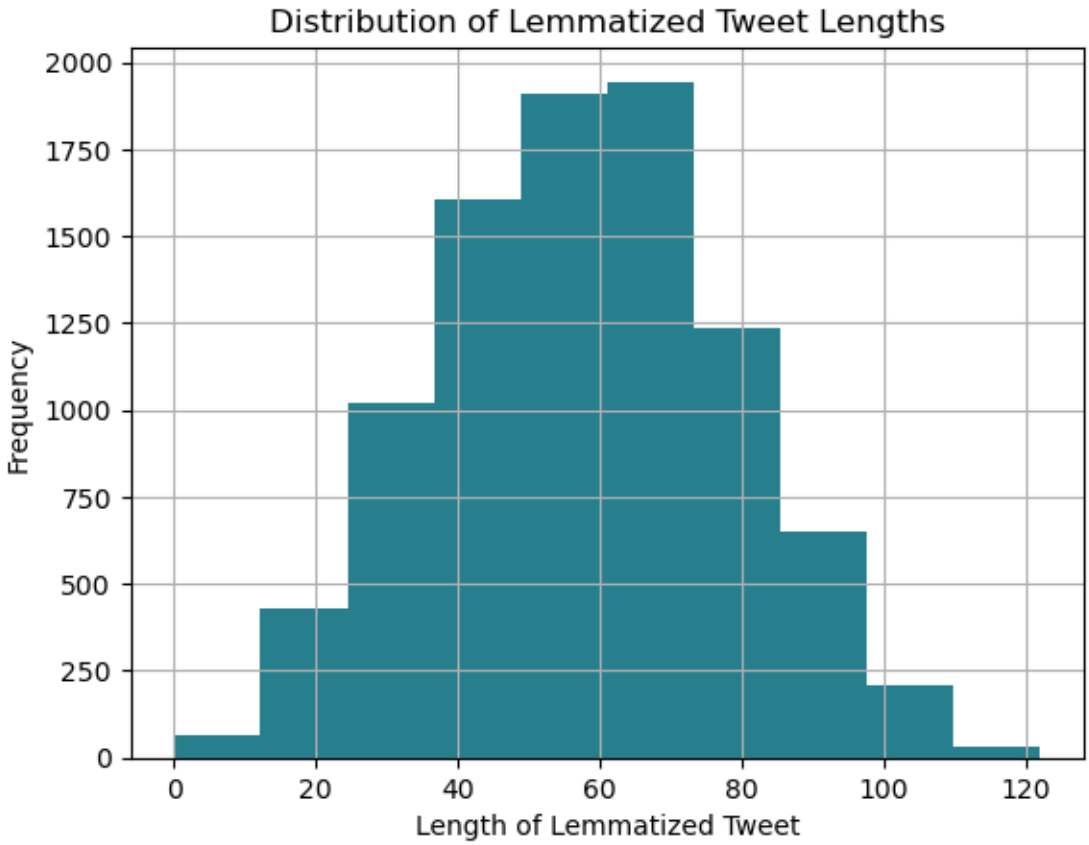
4.1.1: A bar plot of distribution of 'emotion' (target variable)



Observation

- The most frequently occurring emotion is Neutral emotion, with a count of more than 5000, followed by positive emotion with more than 2500 counts. Negative emotions with more than 500 counts and finally I can't tell with below 200 counts.
- There is a noticeable disparity between the counts of No emotion toward brand or product (5338) and the other emotions, with Positive emotion being the next most frequent at 2527, and Negative emotion at 570.
- The high frequency of No emotion toward a brand or product might be attributed to users sharing informational or factual content about the products rather than personal opinions or emotional responses.
- A huge class imbalance between the classes

4.1.2: A histogram visualizing the lemmatized tweet column

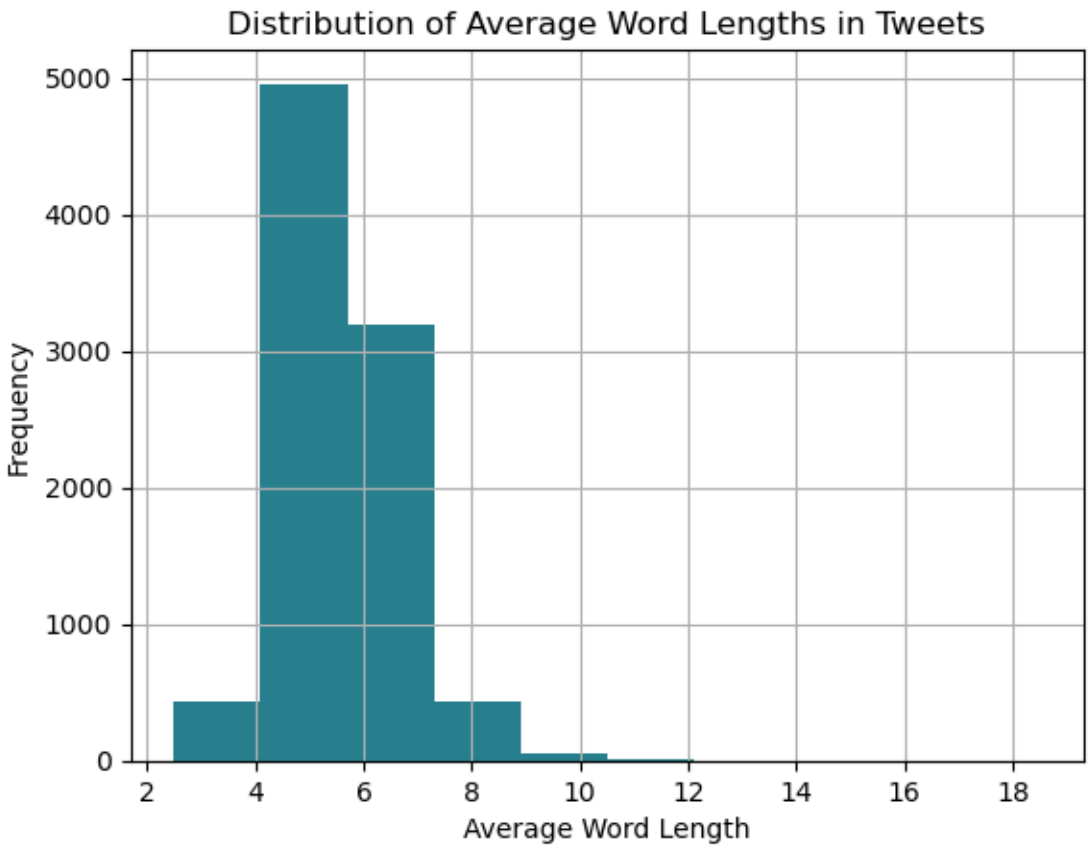


Observations

The histogram shows that news lemmatized tweets range from 10 to 140 characters

There appears to be a normal distribution

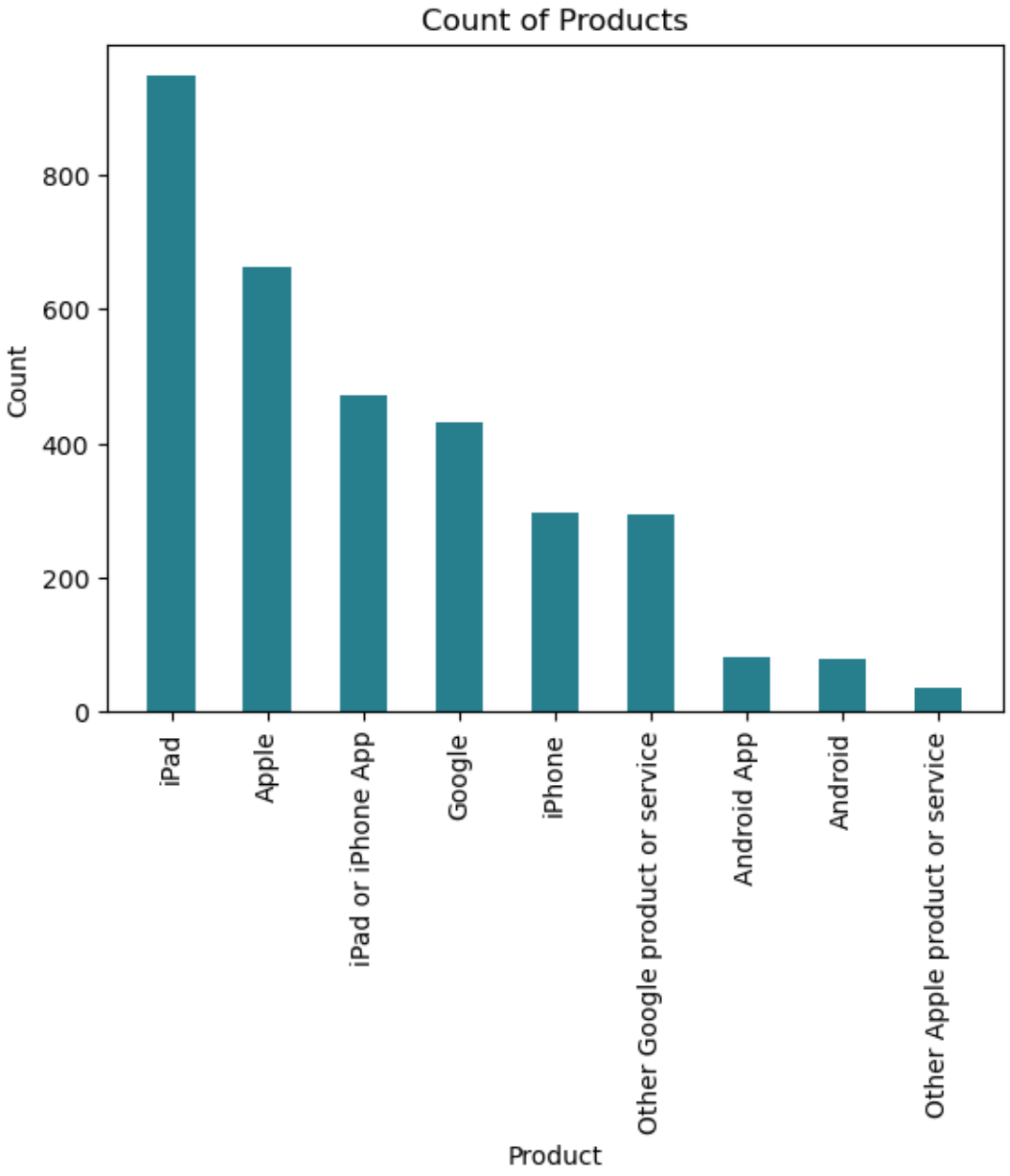
4.1.3: A histogram to visualize the average length of words in Google and Apple products tweets



Observation

The histogram shows that the average word length in these tweets is 5 words, with most tweets falling within the range of 4-5.

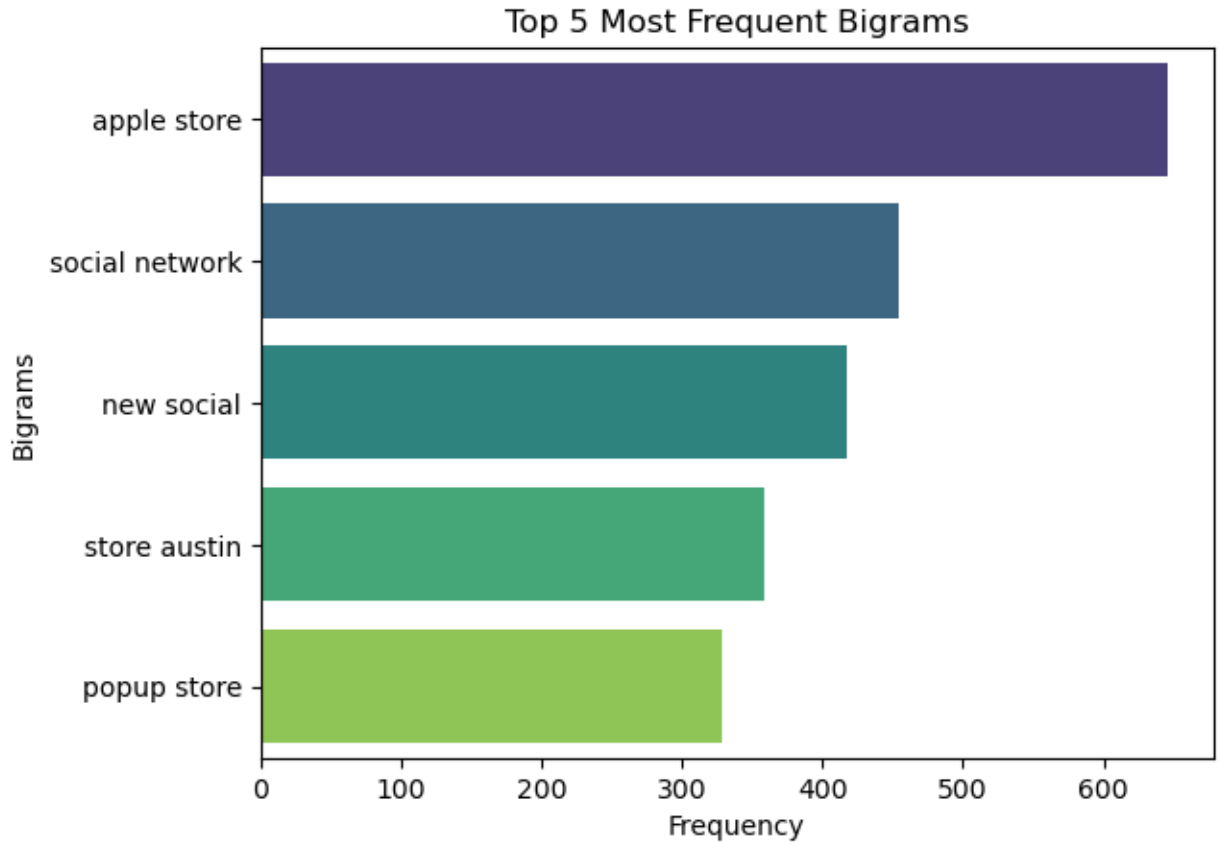
4.1.4: A bar plot showing overview of the products



Observations

The plot above suggests that iPad and apple is the Most Frequently Discussed Products

4.1.5: Most frequent Bargrams used in the dataset

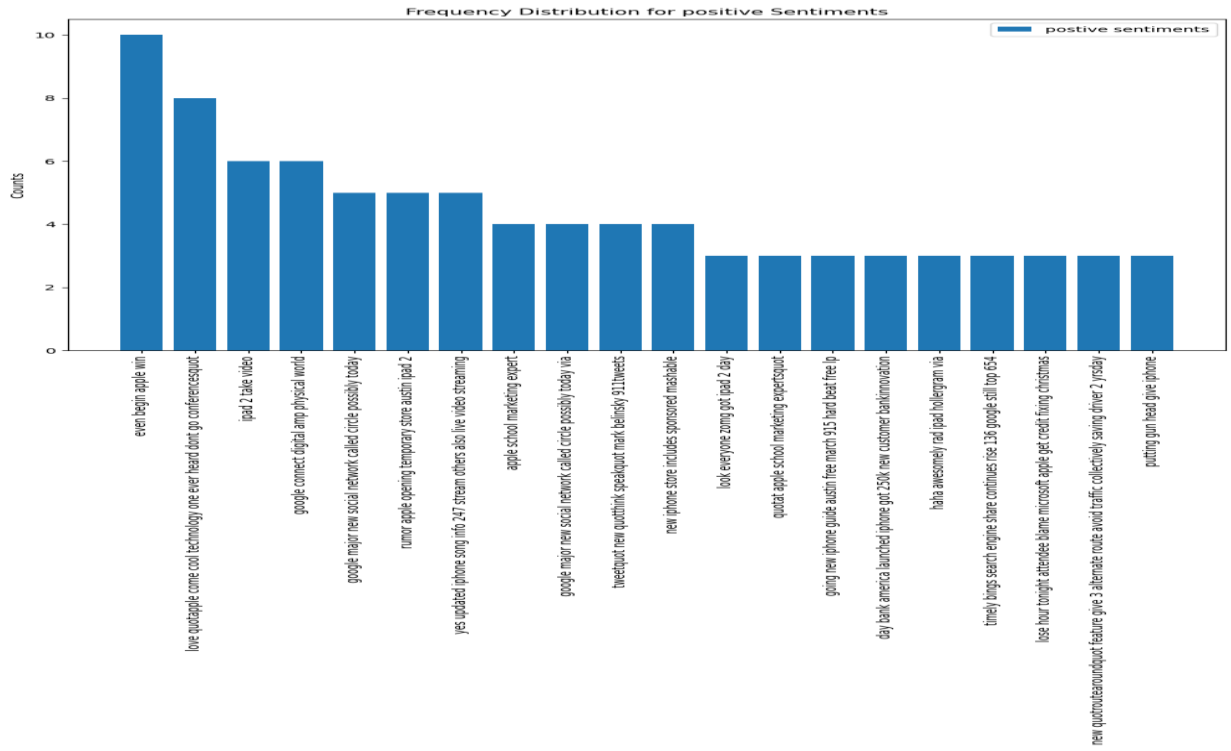


Observation

We can observe that the bigrams such as Apple store are mostly related to dominating the Google and Apple products tweets. Bigrams such as "rt mention" dominating the dataset suggest that many tweets are retweets or mentions, typically used to highlight trending topics, news, and user opinions. This can imply that discussions about Apple and Google products are significantly influenced by social sharing and user interactions on the platform

4.2: Bivalent analysis

4.2.1: A bar plot showing the top positive and negative sentiments



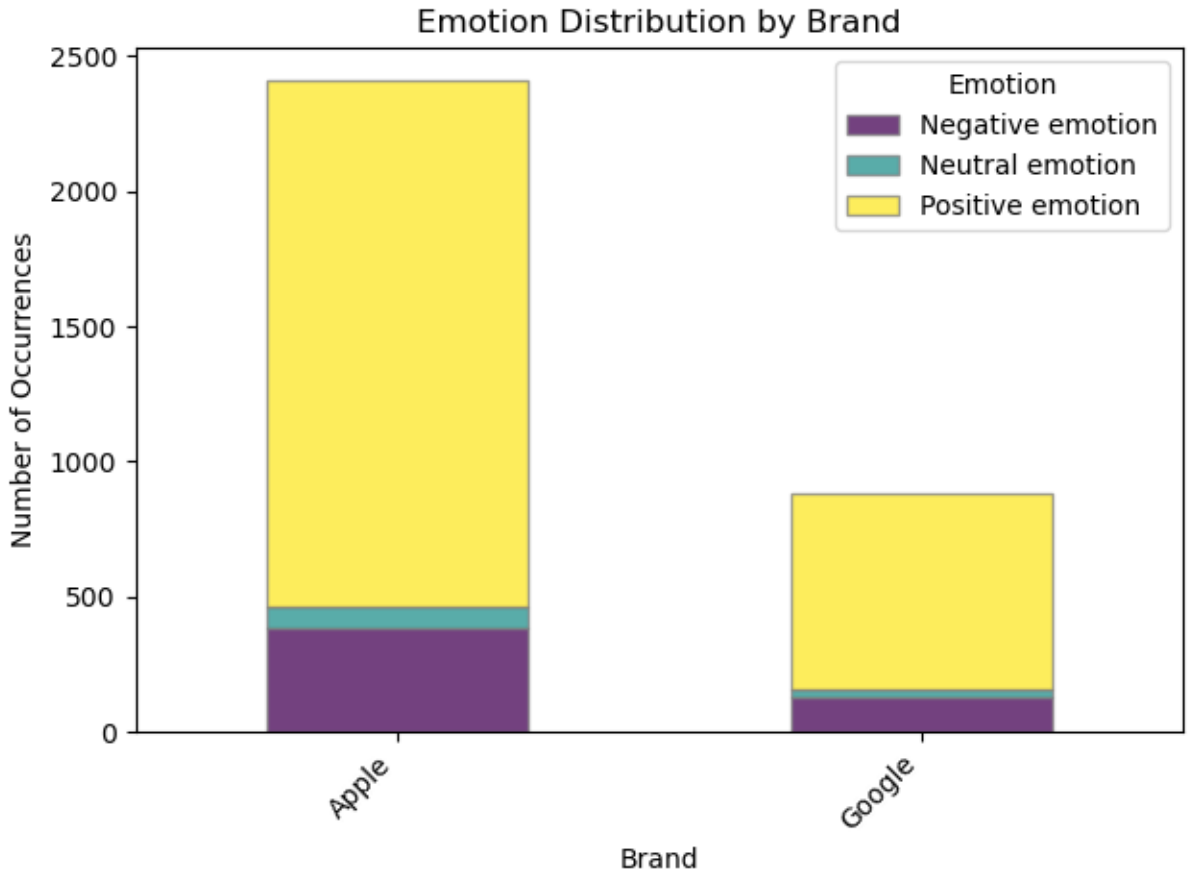
Observation (positive)

Analyzing the most frequent words in positive tweet sentiments reveals interesting insights into customer sentiment toward Apple and Google products. For instance, words like iPad and Google dominate positive sentiments towards Apple, indicating a strong appreciation for it's On the other hand, positive tweets about Google frequently feature words like google This analysis helps google and Apple to understand the specific aspects that drive positive sentiment for each brand, enabling them to tailor marketing strategies and product recommendations accordingly

Observation (negative)

Negative sentiment analysis reveals frequent words like iPad, iPhone, and google indicating customer dissatisfaction with product pricing, performance issues, and potential usability concerns for certain Apple and Google devices.

4.2.2: Visualizing between company name and emotion



Observation

The best brand in regard to positive sentiments is Google while the worst in terms of negative sentiments is Apple

4.2.3: Identifying the most frequent used words used by google and apple products

Observations

The large words in the word cloud provide a snapshot of the key topics and sentiments in our dataset. By focusing on these prominent terms, we can gain insights into user opinions, event influences, and brand mentions, which are crucial for the analysis of sentiment towards Apple and Google products.

The most frequently use words in the google and apple products tweets is google, iPad iPhone, iPad, iPhone

5: Modelling and Evaluation

5.1: Preprocessing

5.1.1: Creating binary and multiclass data frames

We combined values with had either positive or negative emotion together for our binary class and positive, negative and neutral for our multiclass

5.1.2: Label encoding

We used label encoding on our target variable emotion to converting the values into numeric for our models to better understand our target variable

Why label encoding assigns a unique numerical value to each class

5.1.3: Splitting our data into train and test

We identified our independent variable (lemmatized tweets) as our x and emotion as y

And used a test-split size of 80/20. 80% of our data for training and 20% for testing

5.1.4: Creating a pipeline

These pipeline contained a TfidfVectorizer for our lemmatized tweets, smote to correct our class imbalance in the target variable emotion and model as well as its attributes

Why the TfidfVectorizer because it considers words that are frequent in a document but rare overall as more important for that document.

Why use pipeline to prevent data leakage

5.2: Why Binary and Multiclass Classification?

5.2.1: Binary Classification

We decided to perform binary classification to distinguish between 'Positive emotion' and 'Negative emotion.' This task simplifies the problem and focuses on detecting the polarity of sentiments, which is crucial for applications like customer feedback analysis, where identifying positive or negative sentiment can be directly actionable.

5.2.2: Multiclass Classification

We expanded to multiclass classification to include 'Neutral emotion' along with 'Positive emotion' and 'Negative emotion.' This task provides a more nuanced understanding of the sentiments expressed in the tweets, allowing for a more comprehensive sentiment analysis. This is particularly

useful in social media monitoring and understanding public opinion, where not all sentiments are purely positive or negative.

5.3: Binary Classification

5.3.1: Models Used

1. Logistic Regression – **Base Model**
2. Gaussian Naive Bayes
3. Random Forest Classifier

5.3.2: Metrics for Evaluation

- **Accuracy**: Measures the overall correctness of the model.
- **Precision**: The proportion of true positive predictions among all positive predictions. Important for assessing the quality of positive predictions.
- **Recall**: The proportion of true positives correctly identified. Important for understanding the ability of the model to capture all positive instances.
- **F1-Score**: The harmonic mean of precision and recall. Provides a balance between precision and recall.
- **ROC AUC**: Measures the model's ability to distinguish between classes. A higher AUC indicates better performance.

5.3.3: Results

After tuning and evaluating the models, the following results were obtained:

Model	Accuracy	Precision (Positive)	Recall (Positive)	F1-Score (Positive)	ROC AUC
Logistic Regression	0.85	0.93	0.89	0.91	0.87
Gaussian Naive Bayes	0.85	0.91	0.91	0.91	0.83
Random Forest Classifier	0.86	0.89	0.95	0.92	0.82

Best Model for Binary Classification: The **Random Forest Classifier** demonstrated the highest accuracy (86%) and ROC AUC (0.82), making it the best-performing model for the binary classification task. This model balances precision and recall effectively, providing a robust classification performance.

5.4: Multiclass Classification

5.4.1: Models Used

1. Multinomial Naive Bayes – **Base Model**
2. Gradient Boosting Classifier
3. Voting Classifier
4. Stacking Classifier

5.4.2: Metrics for Evaluation

- **Accuracy**: Measures the overall correctness of the model.
- **Precision**: The proportion of true positive predictions among all positive predictions for each class.
- **Recall**: The proportion of true positives correctly identified for each class.
- **F1-Score**: The harmonic mean of precision and recall for each class.
- **ROC AUC**: Measures the model's ability to distinguish between classes for each class. A higher AUC indicates better performance.

5.4.3: Results

After tuning and evaluating the models, the following results were obtained:

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)	ROC AUC (Macro Avg)
Multinomial Naive Bayes	0.64	0.52	0.50	0.51	0.62
Gradient Boosting	0.63	0.57	0.55	0.56	0.68
Voting Classifier	0.68	0.60	0.61	0.60	0.70
Stacking Classifier	0.68	0.60	0.61	0.60	0.70

Best Model for Multiclass Classification: Both the **Voting Classifier** and **Stacking Classifier** showed similar performance, with an accuracy of 68% and a macro average ROC AUC of 0.70. These models balance precision, recall, and F1-score effectively across all classes.

5.5: Rationale for Metric Choices

- **Accuracy**: Provides a straightforward measure of overall model performance but can be misleading in the presence of class imbalance.

- **Precision and Recall**: Important for understanding the trade-off between the number of false positives and false negatives. Critical in applications where either false positives or false negatives carry a high cost.
- **F1-Score**: Balances precision and recall, making it a suitable metric when we need to find a balance between the two.
- **ROC AUC**: Measures the model's ability to distinguish between classes, providing a comprehensive view of model performance across different thresholds.

Conclusions

1. Most of the sentiments show no emotion towards the brand
2. The most mentioned product brand was Apple products
3. The best brand in terms of positive emotion is Google
4. The worst brand in terms of negative emotion is Apple
5. Best Binary classification model is Random forest classifier and in the multiclass is Stacking classifier

Recommendations

- Develop strategies to convert neutral sentiments into positive ones.
- Create targeted campaigns to address specific issues in the negative sentiments.
- Analyze why Apple products dominate mentions and benchmark against Apple.
- Stock google products.
- Investigate negative sentiments toward Apple to discover where the issue is.
- Deploy and constant evaluation of the Random Forest for binary classification and Stacking Classifier for multiclass problems; continuously as well as regular refining of the models
- Conduct regular sentiment analysis to track brand perception changes.