**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**

**Introduction**

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: 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)

* Recall: 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)

* Precision: Precision is 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​)

* F1 score: The F1 score is 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​) )

**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:

o Apple-related products (e.g., iPad, iPhone) to Apple.

o Google-related products to Google.

o 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:**

o All tweets were converted to lowercase for uniformity.

o A list of tweets was created for tokenization.

2.   **Tokenization:**

o A tokenizer specifically designed for Twitter, TweetTokenizer was used to dissect tweets into tokens.

3.  **Hashtag and Accent Removal:**

o Hashtags and accents were removed from tokens to focus on meaningful words.

b.     **Removing Noise**

1.  **Punctuation Removal:**

o Punctuation marks were eliminated from the tokens to reduce noise in the text.

2.   **Stopword Removal:**

o Common stop words were removed from the tokenized text.

o 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."

o 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

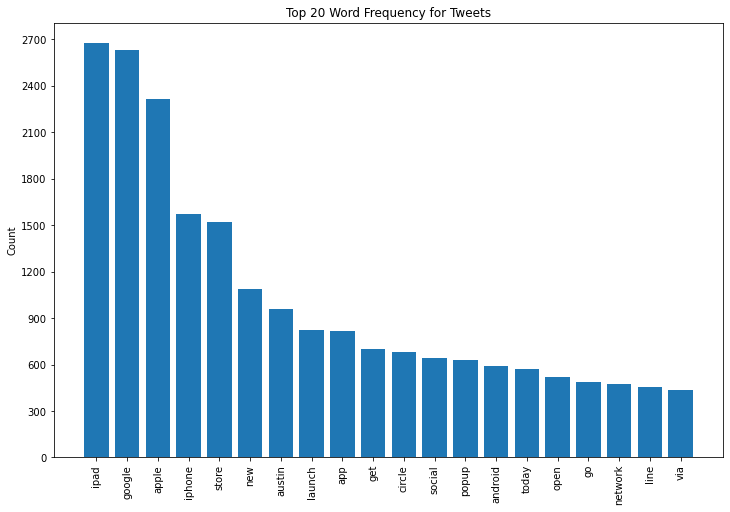
·         Elimination of punctuation

·         Mapping of NLTK POS tags to WordNet POS tags

·         Lemmatization with POS tagging

·         Removal of stop words, irrelevant terms, and special characters

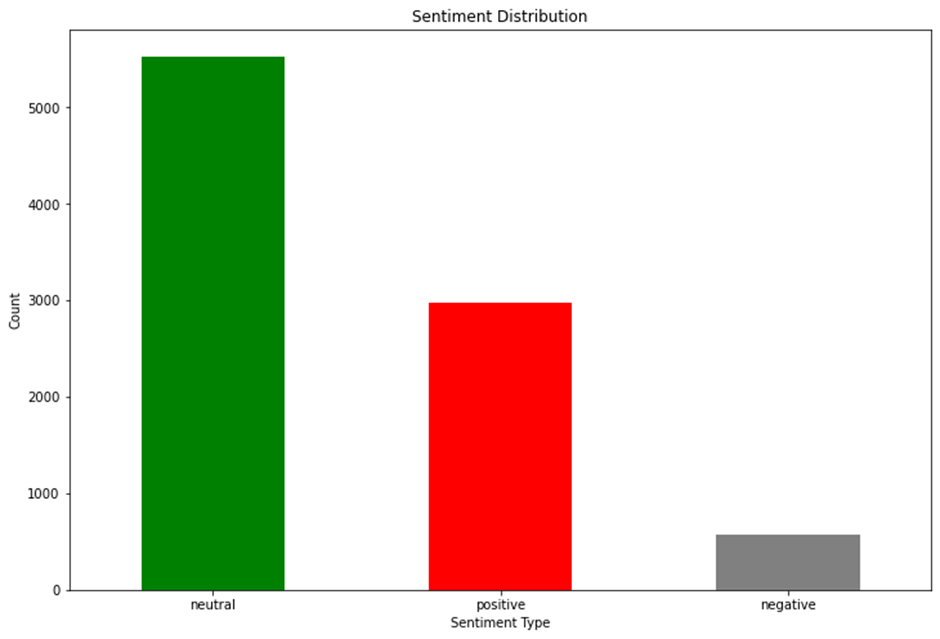
The resulting top 20 words are as shown in the figure below:



The 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 robust foundation for insightful analysis and modelling.

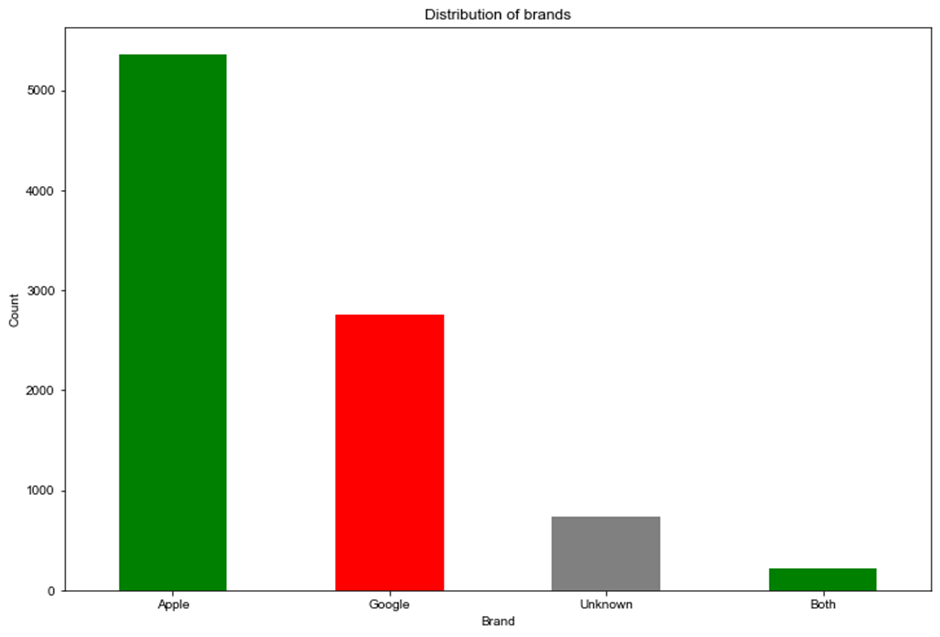
**Exploratory Data Analysis**

**Sentiment Distribution**

****

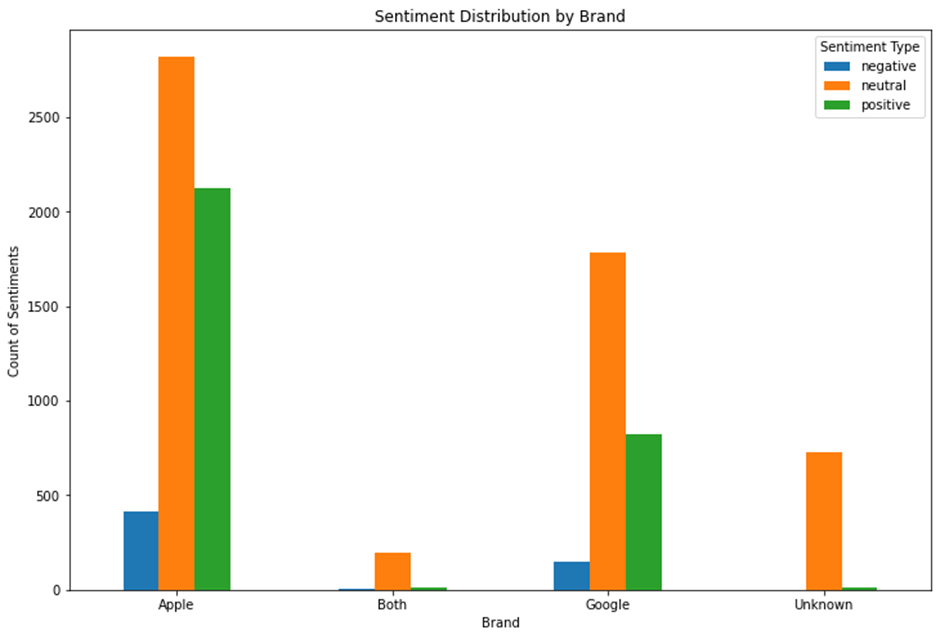
The sentiment distribution graph reveals the relative prevalence of sentiment types across the dataset. Neutral sentiments dominate, suggesting that most tweets neither express strong positivity nor negativity. Positive sentiments come in second, indicating a significant number of favourable opinions. Negative sentiments are the least common, implying that fewer users expressed dissatisfaction or unfavourable opinions.

**Distribution of brands**

****

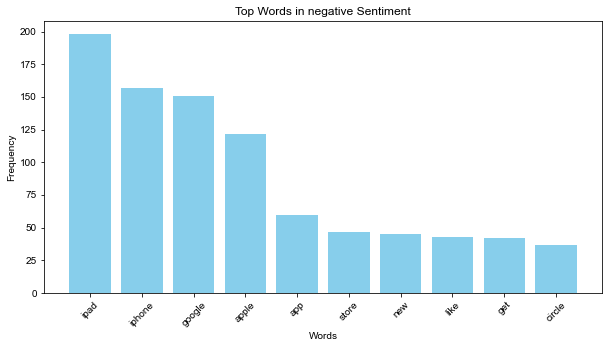
The brand distribution graph illustrates the number of tweets mentioning specific brands. Apple received the highest attention, with over 5,000 tweets, highlighting its prominence during the analyzed period. Google followed with approximately 3,000 tweets. Interestingly, tweets that did not explicitly 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.

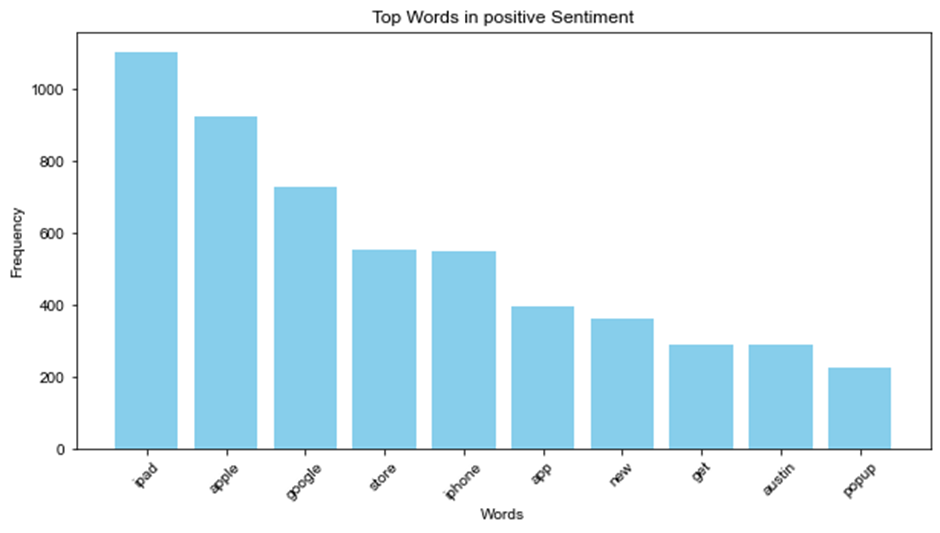
**Sentiment Distribution by Brand**

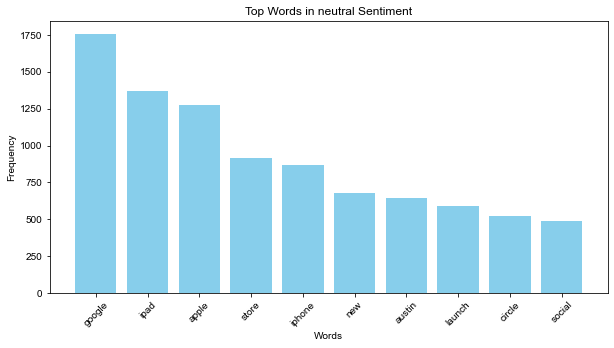
****

This visualization 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. 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.

These analyses collectively offer a comprehensive view of user sentiments and brand engagement, providing valuable insights for tailoring marketing strategies and enhancing product experiences.