**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.

1. **Handling Missing Values and Duplicates**
2. **Checking for Missing Values and Duplicates**

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

1. **Handling Duplicates**

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

1. **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.

1. **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.

1. **Text Data Preparation and Analysis**

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

1. **Initial Preprocessing**
2. **Standardizing Text:**
   * All tweets were converted to lowercase for uniformity.
   * A list of tweets was created for tokenization.
3. **Tokenization:**
   * A tokenizer specifically designed for Twitter, TweetTokenizer was used to dissect tweets into tokens.
4. **Hashtag and Accent Removal:**
   * Hashtags and accents were removed from tokens to focus on meaningful words.
5. **Removing Noise**
6. **Punctuation Removal:**
   * Punctuation marks were eliminated from the tokens to reduce noise in the text.
7. **Stopword Removal:**
   * 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.
8. **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.

1. **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."

1. **Automating the Process**

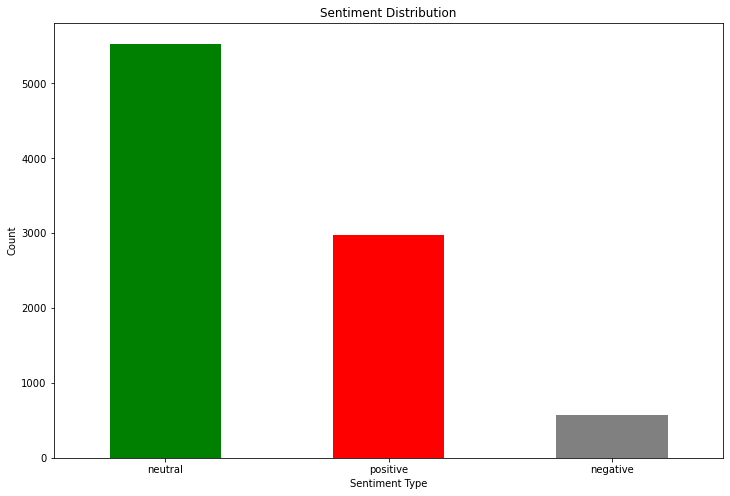
A function was created to streamline the entire text preprocessing workflow for the Tweet column. The steps included:

* 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 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**

A green red and grey rectangles

Description automatically generated

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**

A graph of different colored squares

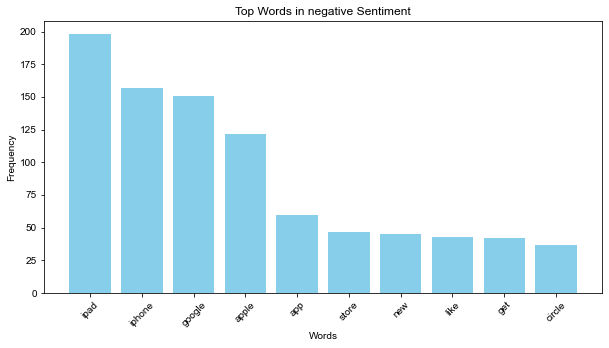
Description automatically generated with medium confidence

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**

A blue and white graph

Description automatically generated



A blue and white graph

Description automatically generated

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.