2024

SentimentFlow

A NON-TECHNICAL PRESENTATION ON SENTIMENT ANALYSIS USING NATURAL LANGUAGE PREPROCESSING

1.0. Business Understanding

1.1. Overview

The world today is rife with information flowing from millions of users across different platforms based on a variety of topics including politics, celebrities, data science, and exercise to make my brain bigger. These opinions on the web garner more and more traffic and gain traction. At the same time, this information reaches a much larger audience who may also share the same information with their networks.

A technique known as Sentiment Analysis tackles the analysis of sentiments using Natural Language Processing.

Natural Language Processing is a machine learning technology that gives computers the ability to interpret, manipulate, and comprehend human language. This would be very useful in analysing human made opinions on the web.

These sentiments across the internet can be analysed using Natural Language Processing methodologies.

Every company/ business with an online presence, and even ones without, require some form of observing, recording, tracking and analysing of these online opinions of their products or services. Doing so will insure their business' public image and ensure that opinions on the web do not burn the palettes of their users, and especially those of the potential users of their products or services, so to speak.

SentimentFlow leverages the power of cutting-edge NLP techniques to analyze sentiment in textual data, providing valuable insights for decision-making by the management of the vendor. This analysis would determine whether sentiments are positive, negative or neutral.

1.2. Problem Statement

With such a large volume of information shared by and / or received from many users and potential users, business would not be able to keep up with the information received if they attempt to track everything, everywhere all at once, manually.

Without fully comprehending the effects of the publics' opinion, the businesses' public image would be tarnished. The poor public image could lead to potentially market share losses, loss of trust from it's repeat consumers, low credibility to its potential clients and also loss of investment/ partnership opportunities.

1.3. Stakeholders

1. **Companies (Apple and Google):** These organizations are directly impacted by public sentiment. They want to monitor how their products are perceived and identify areas for improvement.

2. **Marketing Teams:** Marketing teams can use sentiment analysis to adjust their campaigns, respond to negative feedback, and highlight positive aspects of their products.

3. **Decision-Makers:** Executives and managers need insights into public sentiment to make informed decisions about product development, customer support, and brand reputation.

1.4. Proposed Solution

Analysing the public opinion would help businesses monitor their brand and sentiments around their products and services coming in as customer feedback, and understand customer needs, while making them more conscious thus preventing poor public relations.

1.5. Value Proposition

By accurately classifying tweets, our NLP model can provide actionable insights to stakeholders. For example:

- Identifying negative sentiment can help companies address issues promptly.
- Recognizing positive sentiment can guide marketing efforts and reinforce successful strategies.
- Understanding neutral sentiment can provide context and balance.

1.6. Objectives

Main Objective
To create a NLP multiclass classification model that can analyse sentiments in either 3 categories - Positive, Negative or Neutral. This model targets to achieve a recall score of 80% and an accuracy score of 80%.

Specific Objectives

- To identify the most common words used in the dataset using Word cloud.
- To confirm the most used words that are positively and negatively tagged.
- To recognize the products that have been opined by the users.
- To spot the distribution of the sentiments.
- To develop market strategy that improves the product positioning.

2.0. Data Understanding

Data Sources

The dataset originates from CrowdFlower via data.world. Contributors evaluated tweets related to various brands and products. Specifically:

- Each tweet was labeled as expressing positive, negative, or no emotion toward a brand or product.

- If emotion was expressed, contributors specified which brand or product was the target.

Suitability of Data

Here's why this dataset is suitable for our project:

1. **Relevance:** The data directly aligns with our business problem of understanding Twitter sentiment for Apple and Google products.

2. **Real-World Context:** The tweets represent actual user opinions, making the problem relevant in practice.

3. **Multiclass Labels:** We can build both binary (positive/negative) and multiclass (positive/negative/neutral) classifiers using this data.

Dataset Size

The dataset contains over 9,000 labeled tweets. We'll explore its features to gain insights.

Descriptive Statistics

- **tweet_text:** The content of each tweet.

- **is_there_an_emotion_directed_at_a_brand_or_product:** No emotion toward brand or product, Positive emotion, Negative emotion, I can't tell

- **emotion_in_tweet_is_directed_at:** The brand or product mentioned in the tweet.

Feature Inclusion

Tweet text is the primary feature. The emotion label and target brand/product are essential for classification.

Limitations

- **Label Noise:** Human raters' subjectivity may introduce noise.

- **Imbalanced Classes:** We'll address class imbalance during modeling.
- **Contextual Challenges:** Tweets are often short and context-dependent.
- **Incomplete & Missing Data:** Could affect the overall performance of the models.

2.2. Data

SHAPE
Records in dataset are 9093 with 3 columns.

COLUMNS
Columns in the dataset are:
- tweet_text
- emotion_in_tweet_is_directed_at
- is_there_an_emotion_directed_at_a_brand_or_product

UNIQUE VALUES
Column *tweet_text* has 9065 unique values

Column *emotion_in_tweet_is_directed_at* has 9 unique values
Top unique values in the *emotion_in_tweet_is_directed_at* include:
- iPad
- Apple
- iPad or iPhone App
- Google
- iPhone
- Other Google product or service
- Android App
- Android
- Other Apple product or service

Column *is_there_an_emotion_directed_at_a_brand_or_product* has 4 unique values
Top unique values in the *is_there_an_emotion_directed_at_a_brand_or_product* include:
- No emotion toward brand or product
- Positive emotion
- Negative emotion
- I can't tell

MISSING VALUES
Column *tweet_text* has 1 missing values.
Column *emotion_in_tweet_is_directed_at* has 5802 missing values.
Column *is_there_an_emotion_directed_at_a_brand_or_product* has 0 missing values.

DUPLICATE VALUES

The dataset has 22 duplicated records.

Conclusions from the Data Understanding:

1. All the columns are in the correct data types.
2. The columns will need to be renamed.
3. Features with missing values should be renamed from NaN.
4. Duplicate records should be dropped.
5. All records with the target as "I can't tell" should be dropped.
6. Corrupted records should be removed.
7. Rename values in the is_there_an_emotion_directed_at_a_brand_or_product where the value is 'No emotion toward brand or product' to 'Neutral Emotion'

3.0. Data Cleaning

From the analysis, the intricate steps followed below cleaned the data before further analysis and modeling.

Validity Checks:

- All corrupted records were removed from the dataset,
- Removed all the sentiments that we would not account for.
- Streamlined the values in the third column.

Completeness Checks:

- Dropped any records with missing values in the first column.
- Filled in the missing values in the second column using signposts found in the tweet column.
- Streamlined the values in the emotions column

Consistency Checks:

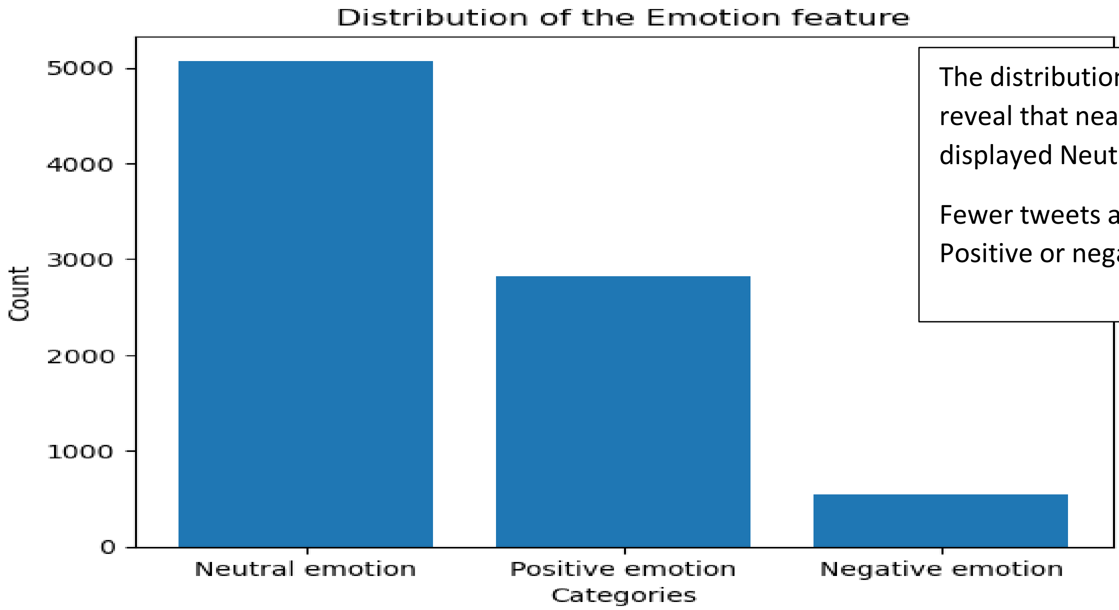
- Dropped any duplicated records in the dataset

Uniformity Checks:

- Renamed the columns
- Reset the index of the data.

4.0. Data Visualization

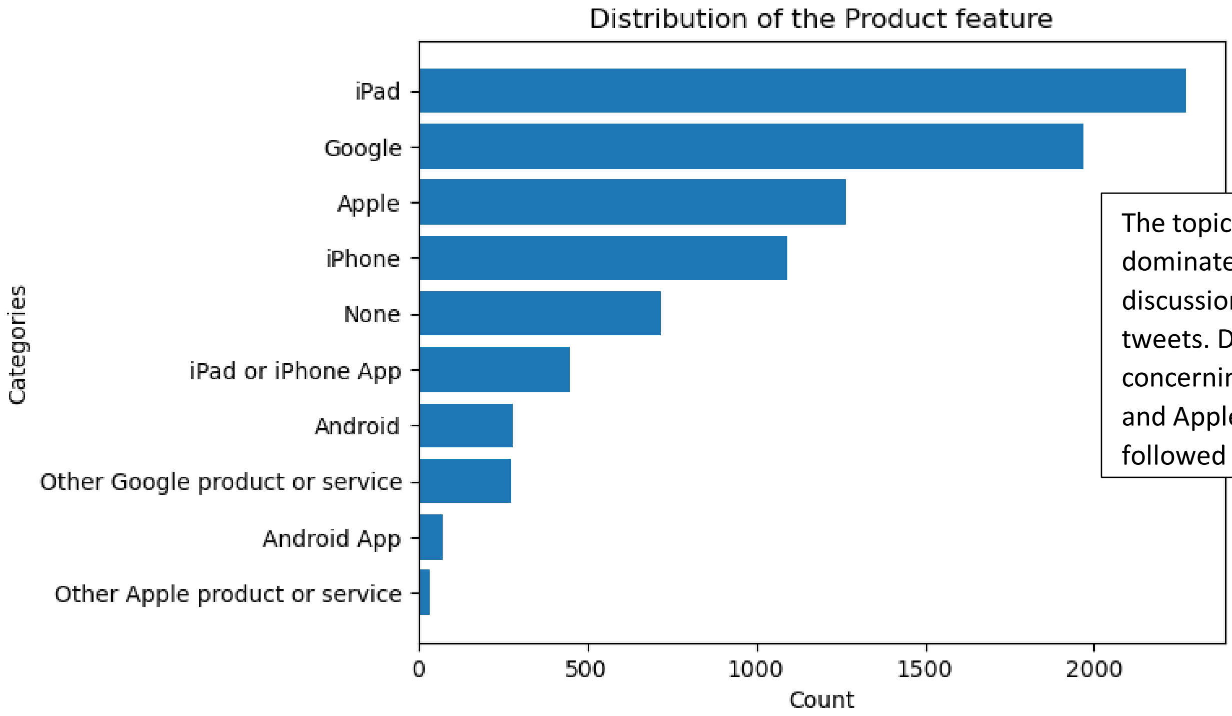
4.1. Distributions of the Emotions Feature



The distribution of the emotions reveal that near 5000 of the tweets displayed Neutral emotions.

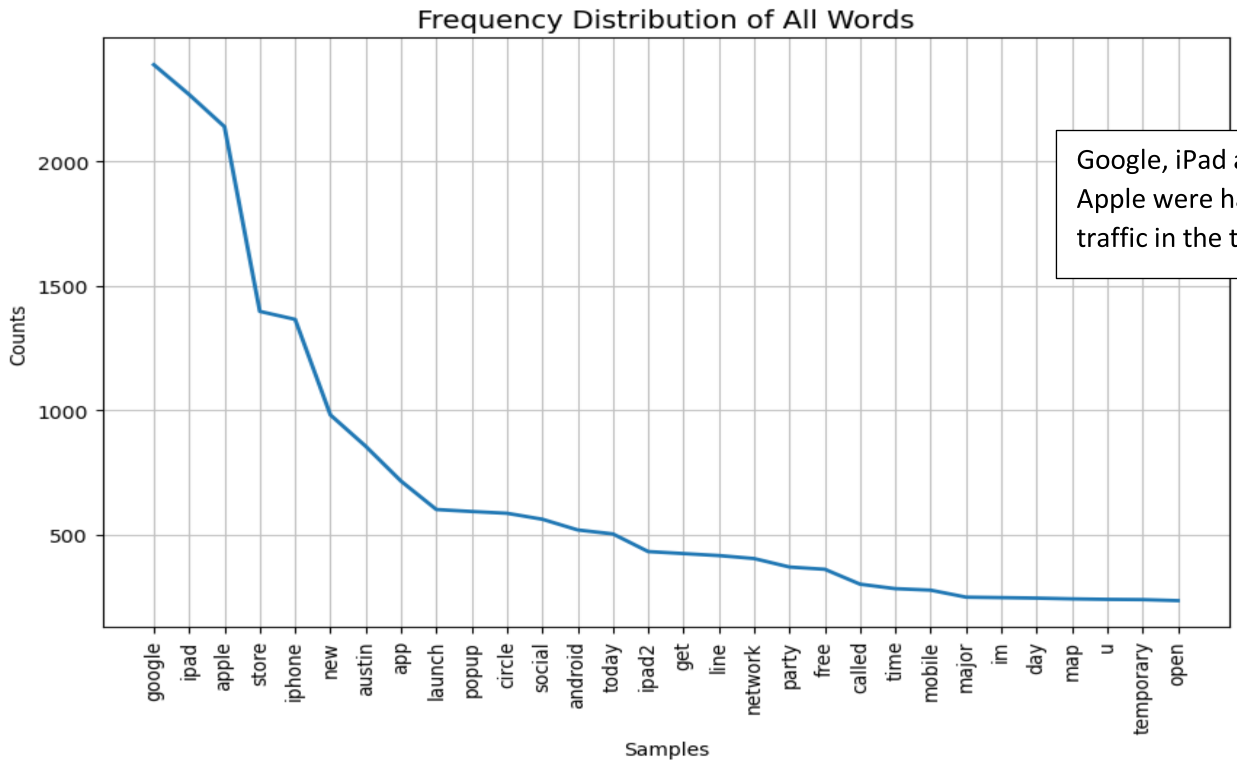
Fewer tweets are categorized as Positive or negative emotions.

4.2. Distributions of the Products Feature

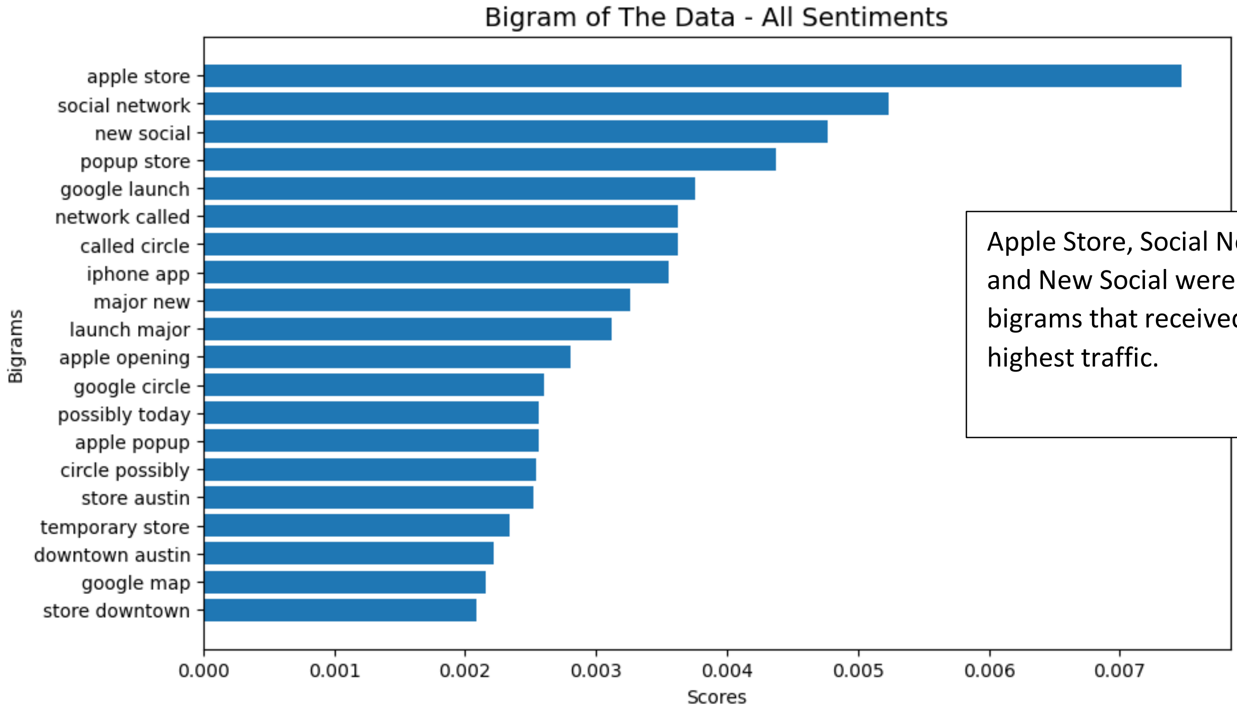


The topic iPad dominated the discussion in the tweets. Discussions concerning Google and Apple companies followed in ranking.

4.3. Word Distribution of all the tweets across sentiments

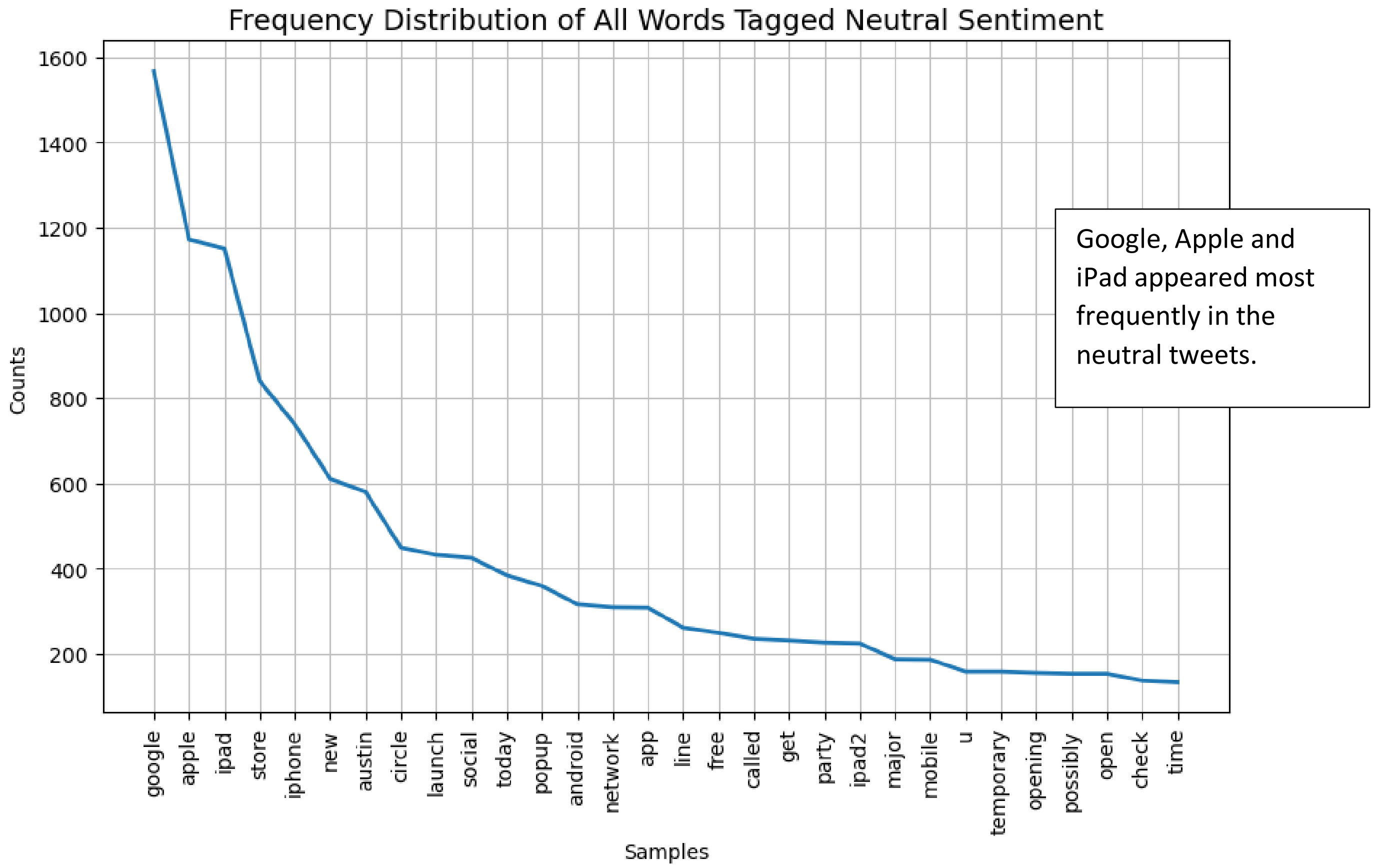


Google, iPad and Apple were had high traffic in the tweets.

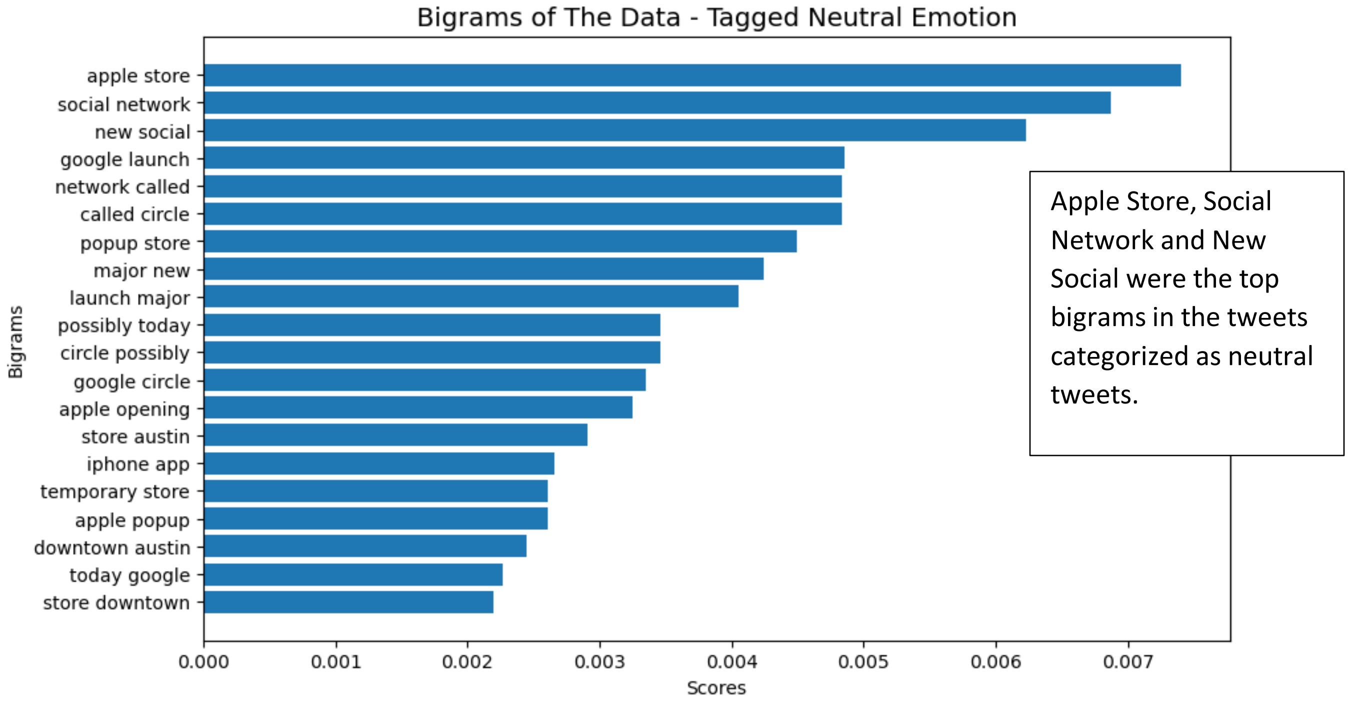


Apple Store, Social Network and New Social were the bigrams that received the highest traffic.

4.4. Distribution of all the tweets classified as Neutral

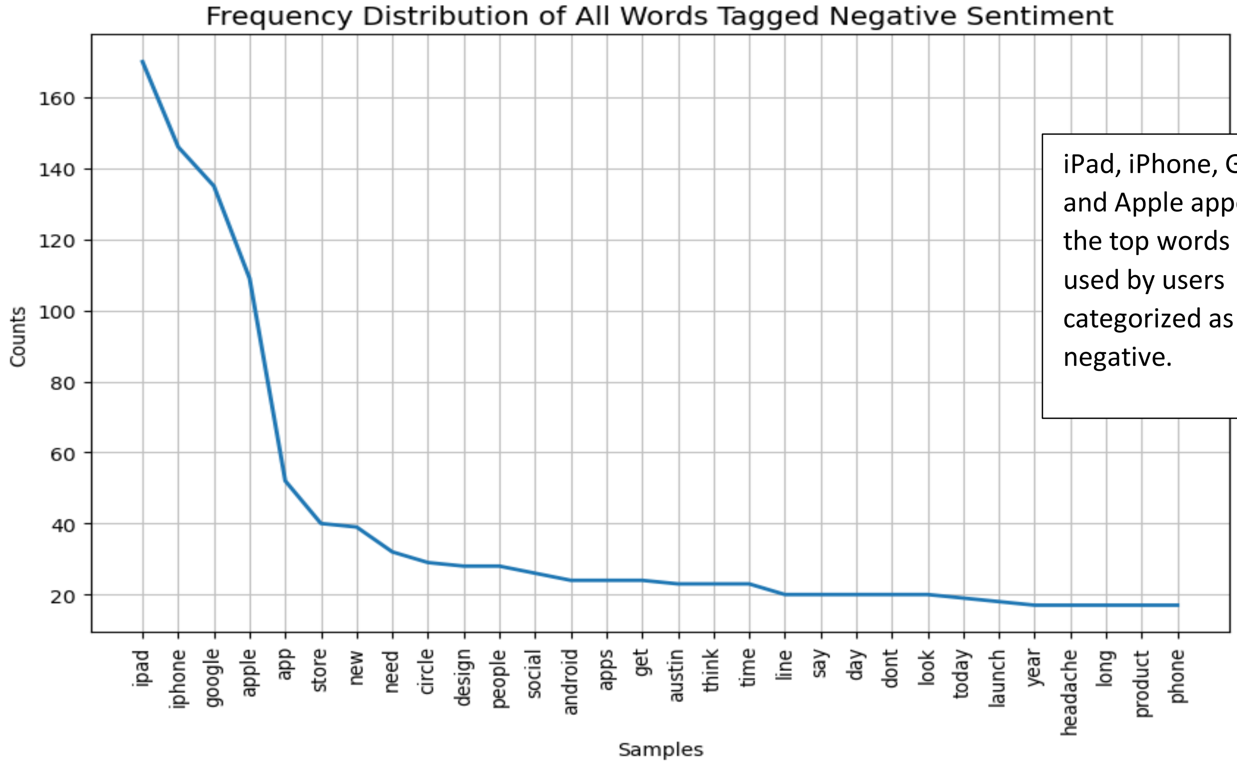


Google, Apple and iPad appeared most frequently in the neutral tweets.

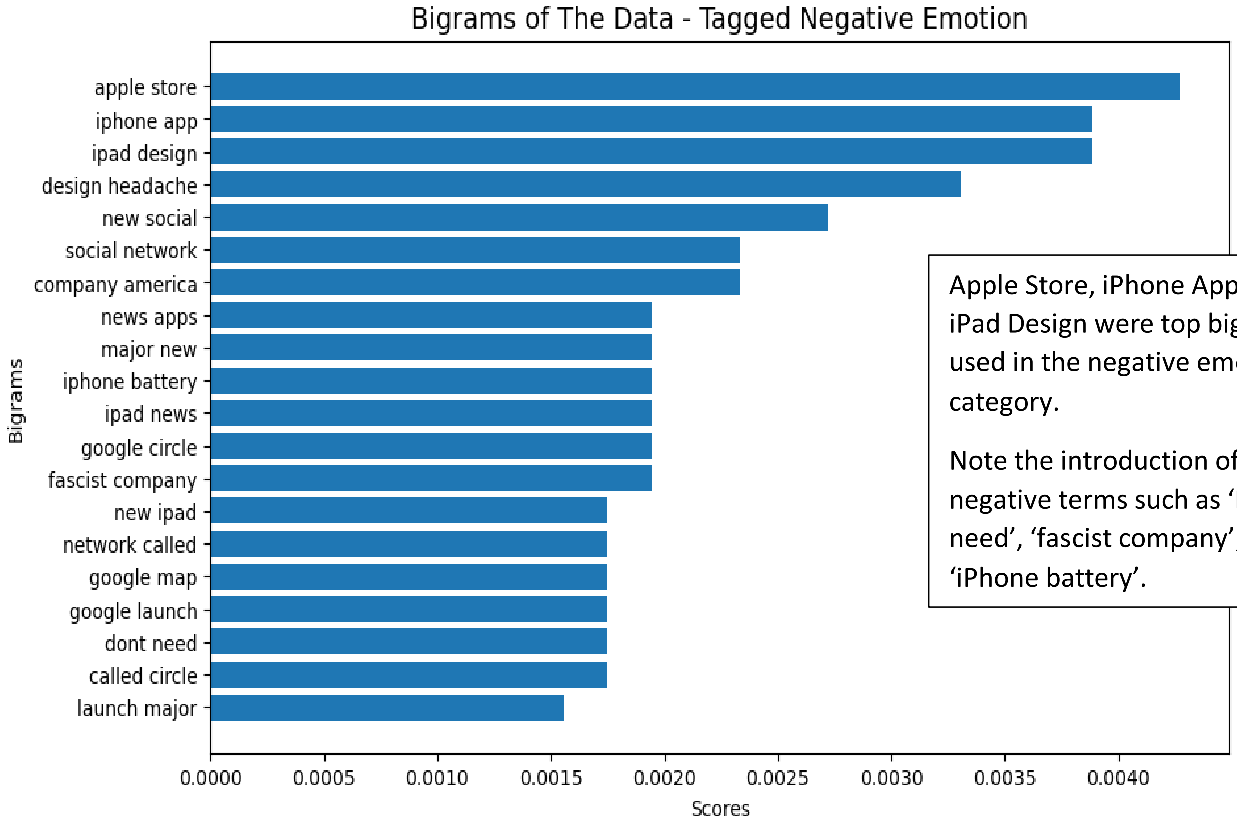


Apple Store, Social Network and New Social were the top bigrams in the tweets categorized as neutral tweets.

4.5. Distribution of all the tweets classified as Negative



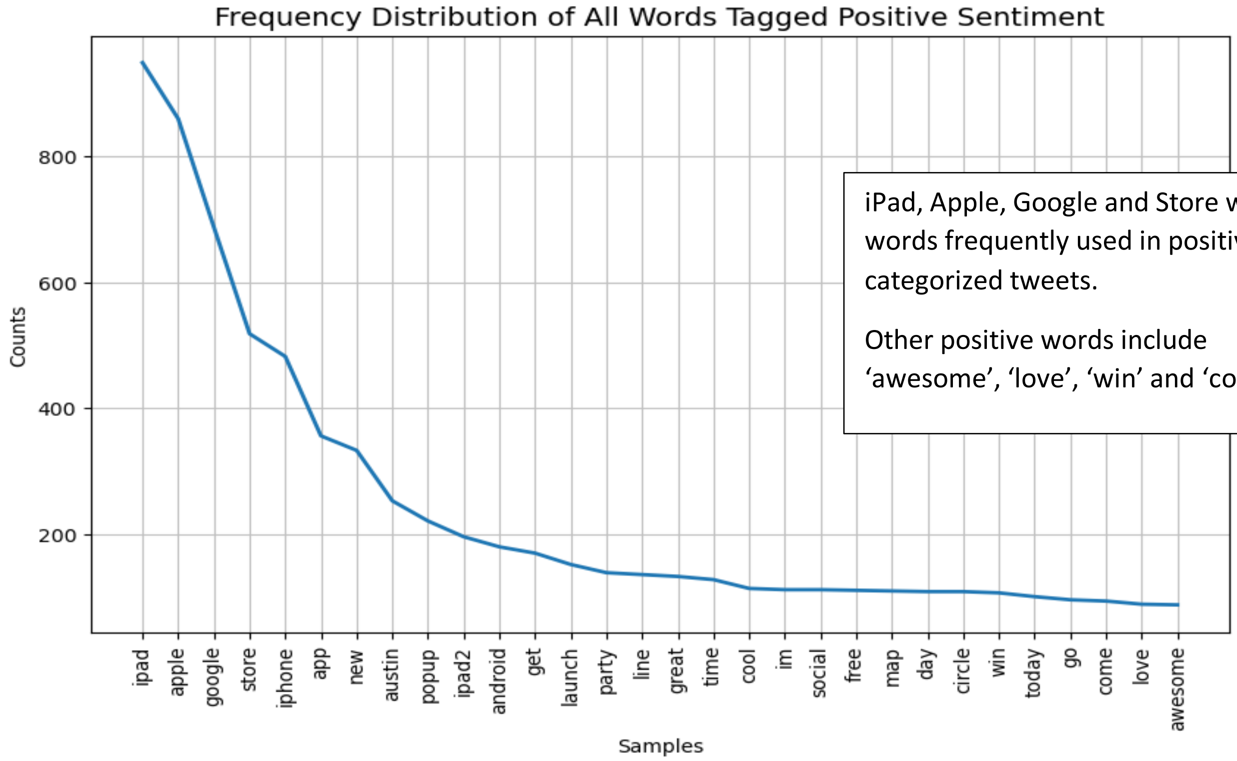
iPad, iPhone, Google and Apple appeared as the top words most used by users categorized as negative.



Apple Store, iPhone App and iPad Design were top bigrams used in the negative emotion category.

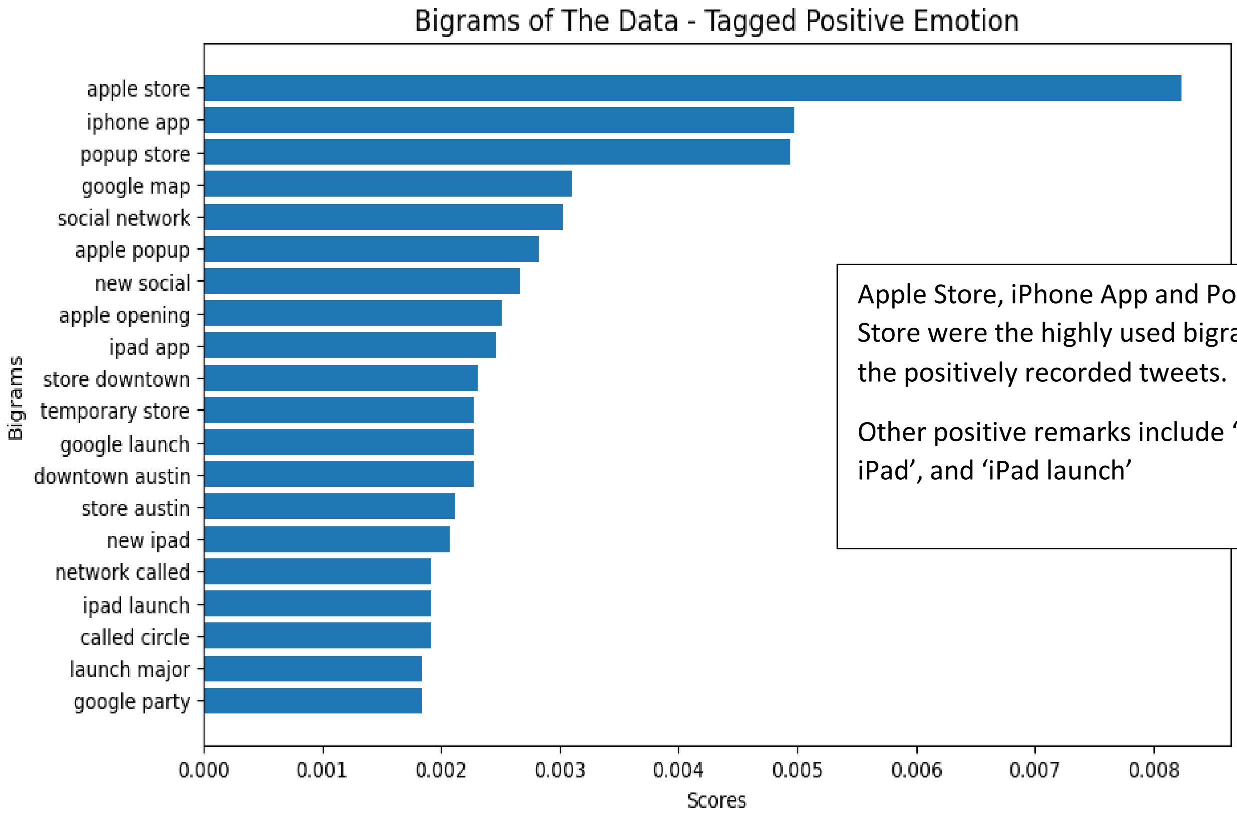
Note the introduction of other negative terms such as ‘Don’t need’, ‘fascist company’, and ‘iPhone battery’.

4.6. Distribution of all the tweets classified as Positive



iPad, Apple, Google and Store were words frequently used in positively categorized tweets.

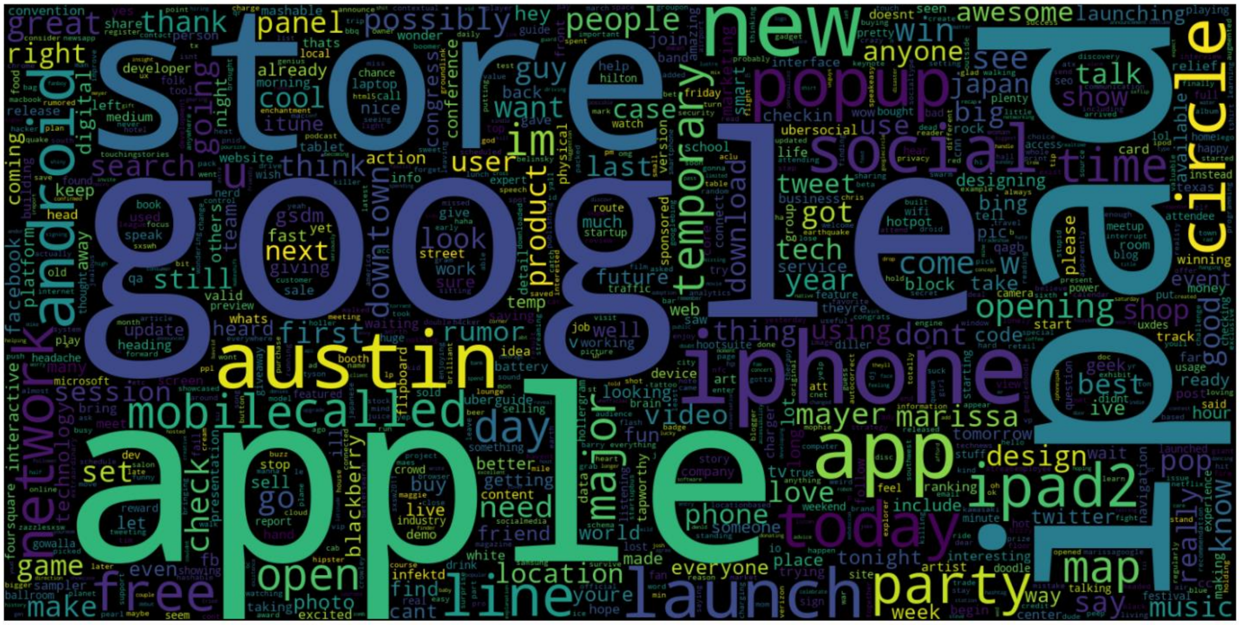
Other positive words include ‘awesome’, ‘love’, ‘win’ and ‘cool’.



Apple Store, iPhone App and Popup Store were the highly used bigrams in the positively recorded tweets.

Other positive remarks include ‘new iPad’, and ‘iPad launch’

4.7. Word Cloud



The image above shows a word clouding strongly showing words such as ‘google’, ‘apple’, ‘iPad’ and ‘store’ appearing as the most used words in the tweets.

5.0 Modeling

Machine Learning involves creating algorithms that learn from data to make predictions or decisions. These algorithms find patterns in data, which are used to predict future trends or classify new data points. In this section, we aim is to evaluate these models to classify tweets into emotions like positive, negative, or neutral

5.1. Overview of Machine Learning Models

Models considered for the project:

1. **Logistic Regression:** This statistical model predicts the probability of a binary outcome. For multi-class classification, it extends to multinomial logistic regression. It works well for problems where the relationship between features and the target variable is approximately linear. We use it to classify tweets into different emotion categories by estimating the probabilities of each class.

2. **Naive Bayes (Multinomial Naïve Bayes):** This model based - on Bayes' theorem with the assumption of independence between features. The Multinomial Naive Bayes variant - is particularly suited for text classification tasks. This model classifies tweets by calculating the likelihood of each emotion given the text features.

3. **Random Forest:** This ensemble learning method builds multiple decision trees and combines their outputs to improve accuracy and prevent overfitting. It aggregates predictions from individual trees to make the final decision. We use Random Forest to leverage its robustness and ability to handle complex datasets with multiple features.

4. **Decision Trees:** This model makes decisions based on asking a series of questions about the features of the data. They split the data into subsets based on feature values, creating a tree-like structure of decisions. Decision Trees classify tweets by making sequential decisions based on the features extracted from the text.

5.2. Data Preprocessing

This is a critical step in the data science and machine learning pipeline that involves preparing and transforming raw data into a format suitable for analysis and modeling. The process ensures that the data is clean, consistent, and structured in a way that enhances the performance of machine learning algorithms. The key objectives of data preprocessing are to improve data quality, handle inconsistencies, and enable more effective model training.

Key Steps in Data Preprocessing:

1. **Data Transformation**: Convert data into a format that is more suitable for analysis and modeling. This often includes normalization, standardization, and encoding. Normalize or standardize numerical features, and encode categorical variables into numerical values (e.g., one-hot encoding, label encoding). We used **label encoding** in this project.

2. **Feature Extraction**: Derive meaningful features from raw data that can improve the performance of machine learning models. Extract features from text or create new features based on existing ones. We used:

- **CountVectorizer**: To convert text into a matrix of token counts.
- **TF-IDF Vectorizer**: To convert text into a matrix of TF-IDF features, which reflect the importance of words in the context of the document and corpus.

3**. Feature Selection**: Identify and select the most relevant features that contribute to the predictive power of the model. Use statistical techniques or algorithms to select features that improve model performance and reduce dimensionality.

4. **Handling Imbalanced Data**: Address class imbalance by generating synthetic examples of minority classes to balance the dataset. We applied the **SMOTE** technique (Synthetic Minority Over-sampling Technique) to balance the dataset and ensure fair model training.

5. Splitting the Dataset: Divide the data into training and testing sets to evaluate model performance, prevent overfitting and are representative of the entire data.

5.3 Model Evaluation

1. Metrics:

- **Accuracy**: Measures the proportion of correctly classified instances out of the total instances.
- **Recall**: Measures the ability of the model to identify all relevant instances of a class.
- **Precision**: Measures the accuracy of the model's positive predictions.
- **F1 Score**: The harmonic mean of precision and recall, providing a single metric for model performance.

2. Hyper-parameter Tuning: Optimize the performance of the models by adjusting hyper-parameters. Grid search or random search to find the best combination of hyper-parameters.

5.4 Results, Interpretation and Insights:

1. Vectorization Performance:

The results underscore the importance of using TF-IDF Vectorization for better feature representation and the benefits of hyper-parameter tuning in optimizing model performance.

2. Model Performance:

Tuned Logistic Regression and **Tuned Random Forest** models demonstrated the highest performance when using TF-IDF Vectorization. Both models achieved approximately **83.7%** accuracy and **83.6%** recall. These metrics indicate their robust capability in classifying tweets accurately and comprehensively.

Multinomial Naive Bayes (Multinomial NB) also performed well, with a final accuracy and recall score of **80%** after tuning.

3. Hyper-parameter Tuning Impact:

The impact of hyper-parameter tuning was notable, with significant improvements in performance metrics. For both Random Forest and Logistic Regression models, accuracy and recall improved by more than 10% due to optimized hyper-parameters.

4. Chosen Model for Deployment:

Tuned Logistic Regression and **Tuned Random Forest** models with **TF-IDF Vectorization** were the most effective models. Given their high accuracy, robust performance, and reliable results, either models is highly recommended for deployment. The choice between them may depend on additional factors such as computational resources or specific deployment requirements.

6.0 Recommendations and Next Steps

6.1. Deployment Recommendations: For real-time sentiment analysis, deploy the Tuned Logistic Regression or Tuned Random Forest models with TF-IDF Vectorization. Both models exhibit high accuracy and robustness. The choice should consider computational resources, response time, and integration requirements with existing systems.

6.2. Practical Applications: The sentiment analysis models can be utilized for various real-world applications, including:

- **Monitoring Social Media:** To track public sentiment about products, brands, or events.
- **Customer Feedback Analysis:** To gain insights from customer reviews and feedback.
- **Market Research:** To identify trends and consumer opinions. Implementing these models can help organizations enhance customer engagement, improve decision-making, and respond to emerging trends effectively.

6.3. Future Work:

- **Data Expansion:** Collect tweets from different geographical areas to understand regional sentiment variations. Scrape more and recent data from X to understand the different sentiments towards the google and apple products. This will provide up to date public opinion of the products.
- **Model Enhancements:** Investigating more sophisticated models or computationally expensive algorithms to improve accuracy and recall.
- **Real-Time Processing Capabilities:** Developing solutions for real-time sentiment analysis to provide timely insights and responses.

7.0 Conclusion

7.1. Summary of Findings: The project successfully evaluated various machine learning models for classifying tweets into emotion categories. TF-IDF Vectorization demonstrated superior performance compared to CountVectorizer, and the Tuned Logistic Regression and Tuned Random Forest models achieved high accuracy and recall. Hyper-parameter tuning significantly enhanced model performance, and the SMOTE technique effectively addressed class imbalance.

7.2. Project Impact: The project's results offer valuable insights into effective sentiment analysis for social media data. The models developed provide robust tools for classifying tweets and understanding public sentiment, contributing to advancements in sentiment analysis and machine learning applications.

7.3. Final Thoughts: The project highlights the importance of effective data preprocessing, model evaluation, and hyper-parameter tuning in achieving high-performance sentiment analysis. The findings emphasize the potential of machine learning models in practical applications and pave the way for future research and enhancements in the field.