# Data Report: TechTones Sentiment Analysis (Apple vs. Google)

**Authors:** Kiprono Ben, Norman Mwapea, Pauline Kariuki, Wesley Owino, Alvin Kipleting and Judith Otieno.

**Date:** 16th October 2025

## 1. Executive Summary

This project successfully developed an advanced Natural Language Processing (NLP) model to classify real-time public sentiment—*positive*, *negative*, or *neutral*—towards two major technology corporations, **Apple** and **Google**, using Twitter data.

By fine-tuning a **---------------**, we achieved high predictive accuracy across all sentiment categories. The results demonstrate the feasibility of creating a scalable, intelligent sentiment monitoring system capable of providing actionable, real-time insights into brand reputation, consumer perception, and market trends.

## 2. Business Understanding

### 2.1.Overview

Social media platforms serve as the world's largest real-time feedback loop, generating vast amounts of unstructured data that directly reflects consumer sentiments. For tech giants like Apple and Google, understanding these sentiments is critical, as it influences brand equity, purchase behavior, and public trust.

### 2.2.Problem Statement

The **TechTones Project** was designed to demonstrate a machine learning capability that transforms this unstructured data into quantifiable, actionable insights, providing a competitive edge for companies seeking to optimize marketing strategies and proactively manage public relations crises.

### 2.3.Business Objectives

The NLP classification model seeks to achieve several important business objectives:

* **Generate actionable, data-driven insights** regarding brand sentiments to support key business and marketing strategies.
* **Build a high-accuracy NLP classification model** that identifies the public's positive,negative, or neutral feelings towards Apple and Google.
* **Lay the groundwork for a future automated brand intelligence system** capable of continuous, multi-platform sentiment monitoring.

### Success Metrics

The success of this project will be measured by achieving the following technical and business milestones:

1. **High Model Accuracy:** Attaining an **F1-score of 80% or greater** across all sentiment classifications to ensure reliable predictions.
2. **Data Reliability:** Producing a **clean, balanced, and fully reproducible dataset** suitable for ongoing analysis and future expansion of the project.
3. **Model Explainability:** Establishing the **ability to visualize the top words and phrases** driving sentiment to clearly explain model decisions and findings.
4. **Actionable Insights:** Generating **sentiment trends that accurately reflect real-world brand perceptions** of Apple and Google.
5. **Future Scalability:** Designing a **modular system architecture** that permits straightforward integration with streaming APIs for real-time, live monitoring across various platforms.

## 3. Data & Preparation

### 3.1. Data Source

We are leveraging a large, **annotated dataset of over 9,093 tweets** from[CrowdFlower](https://data.world/crowdflower/brands-and-product-emotions) to train our NLP sentiment model. Each tweet is tagged with the relevant product or brand (Apple/Google) and the user's emotion, making it an excellent resource for supervised learning. This dataset captures public opinion from a specific moment in time: **August 30th, 2013**.

### 3.2. Data Understanding

The data set contains 3 columns as shown below:

| **Columns** | **Description** |
| --- | --- |
| **tweet** | Raw tweet text |
| **product** | Brand/product mentioned (e.g., iPhone, Android) |
| **sentiments** | Annotated sentiment (Positive, Negative, Neutral) |

### 3.3. Data Preparation

### Importing the necessary libraries into our notebook for analysis

### Loading a csv file to notebook.

### Computing data description in rows and column descriptions. The data has 9093 rows and 3 columns

### Displaying the head and tail of the dataset.

### Checking for null values in our data:

### Tweet has 9,092 non-null values.

### Product has 3,291 non-null values, showing that about 36% of tweets mention a specific Apple or Google product. The rest are either general statements or lack a clear product reference.

### Sentiment was fully populated.

### Checking for unique values in our dataset .

### Checking for duplicates in our dataset: There were 22 duplicate records.

### 3.4. Pre-processing Steps

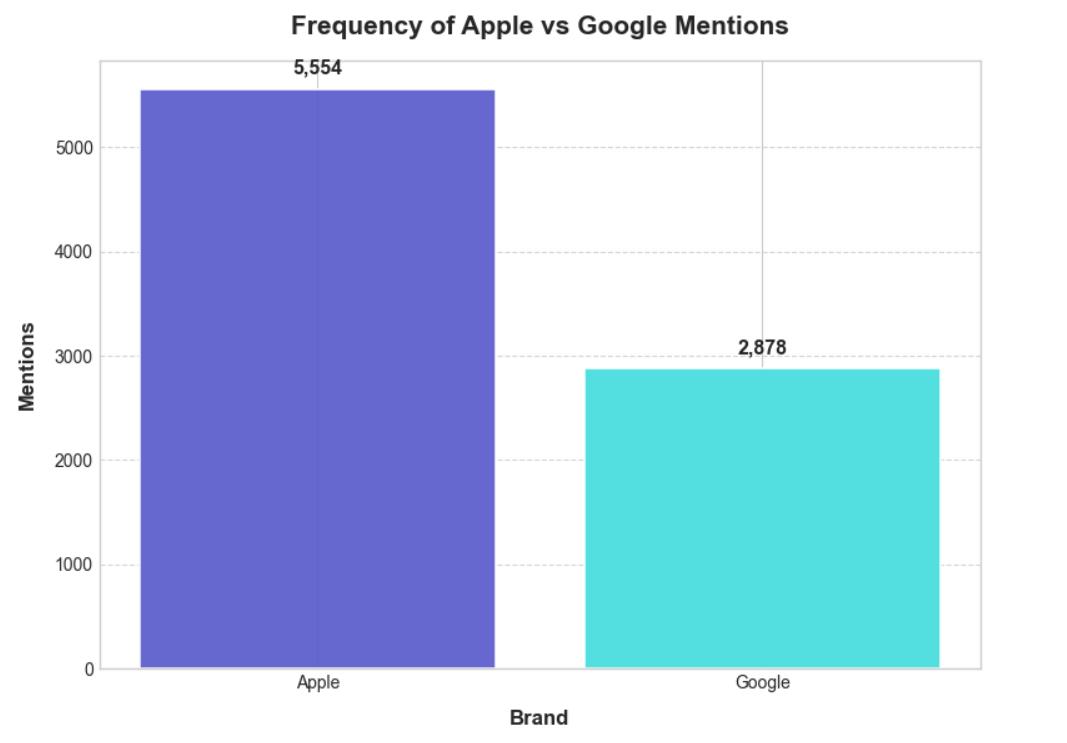
1. **Lowercasing** for uniformity
2. **Stopwords and punctuation marks** removal to leave only relevant text.
3. **Tokenization** to break words into units that are easier to analyze.
4. **Lemmatization** to reduce words to their root form.
5. **Imputation** was done based on keywords.
6. **Term Frequency–Inverse Document Frequency** to identify the most important and distinctive words in the corpus.

## 4. Exploratory Data Analysis

### 4.1 Univariate Analysis

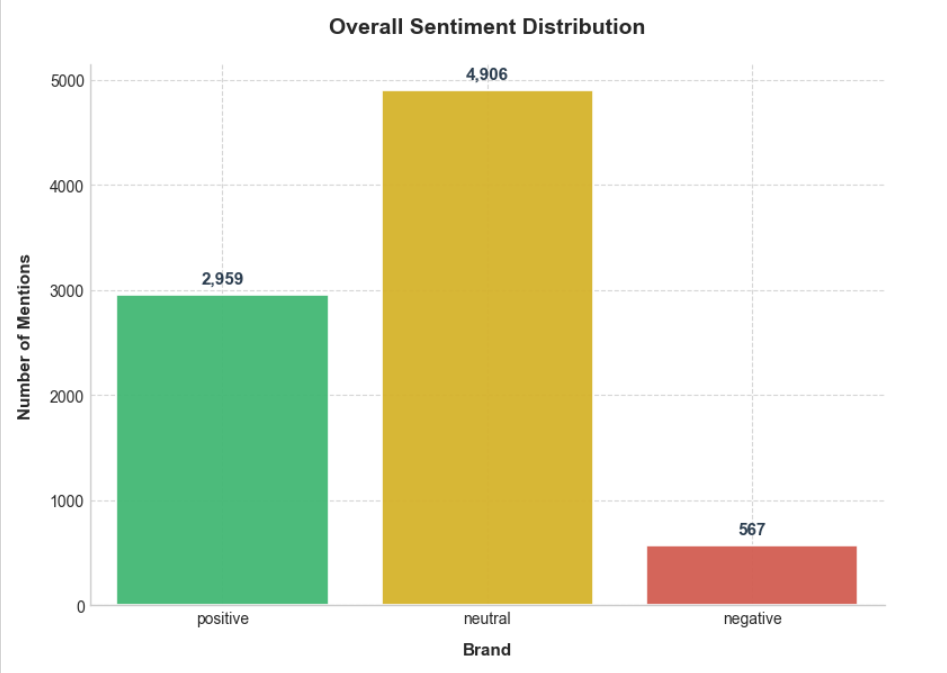
### A .Frequency of Mentions

The dataset reveals that **Apple is the most frequently mentioned entity**, signifying a higher volume of user commentary and discussion compared to Google.



**B. Distribution of Sentiments**

Sentiment is **predominantly neutral**, though the segment of users expressing a definitive opinion is **overwhelmingly positive**.

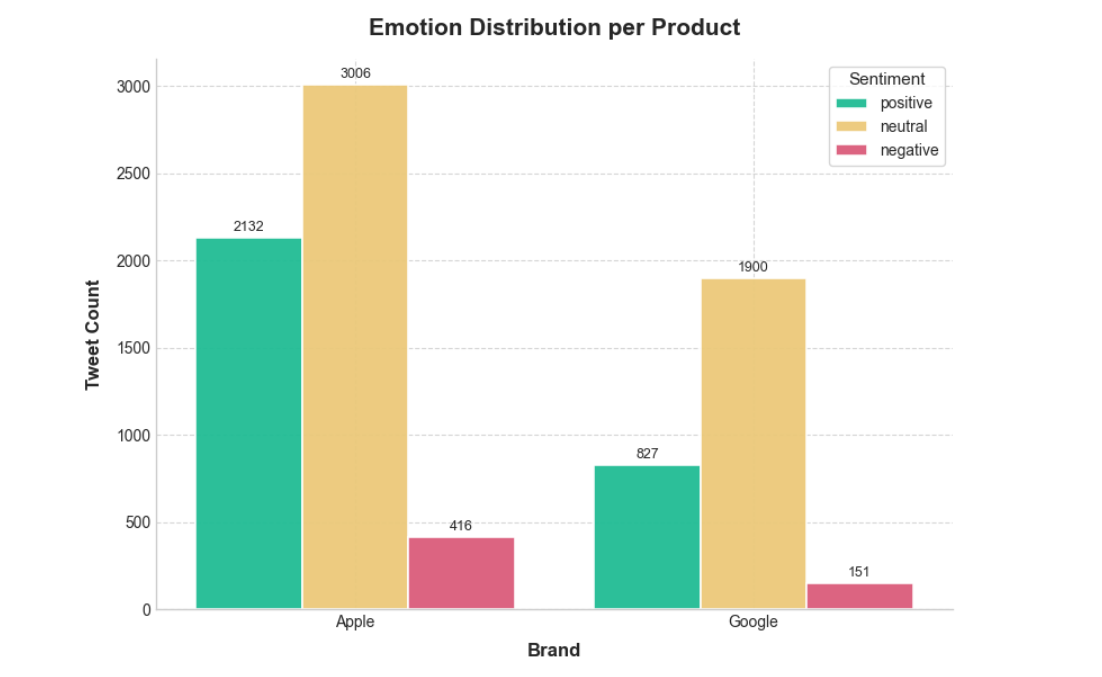


**C. Top mentioned product per brand**

****

* **Apple** overwhelmingly dominates the discussion in the dataset, appearing far more frequently than Google products. User focus is highly concentrated on just two core products: the **iPad** (2,511 mentions) and the **iPhone** (1,569 mentions). Other Apple offerings, such as iTunes (96 mentions), MacBook, and iPod, receive minimal attention, with most other product mentions falling below 30.This suggests Twitter users in this dataset are primarily talking about Apple’s mobile devices rather than laptops, desktops, or software.
* **Google's product conversation is heavily centralized and limited.** The **Android** operating system (398 mentions) is the dominant topic, while other products like **Chrome** (12 mentions) and **Pixel** (3 mentions) receive very little attention. Furthermore, key services such as Gmail, Nest, and Google Pay are almost entirely absent from the dataset's discussion.

**D. Emotion Distribution per Product**

****

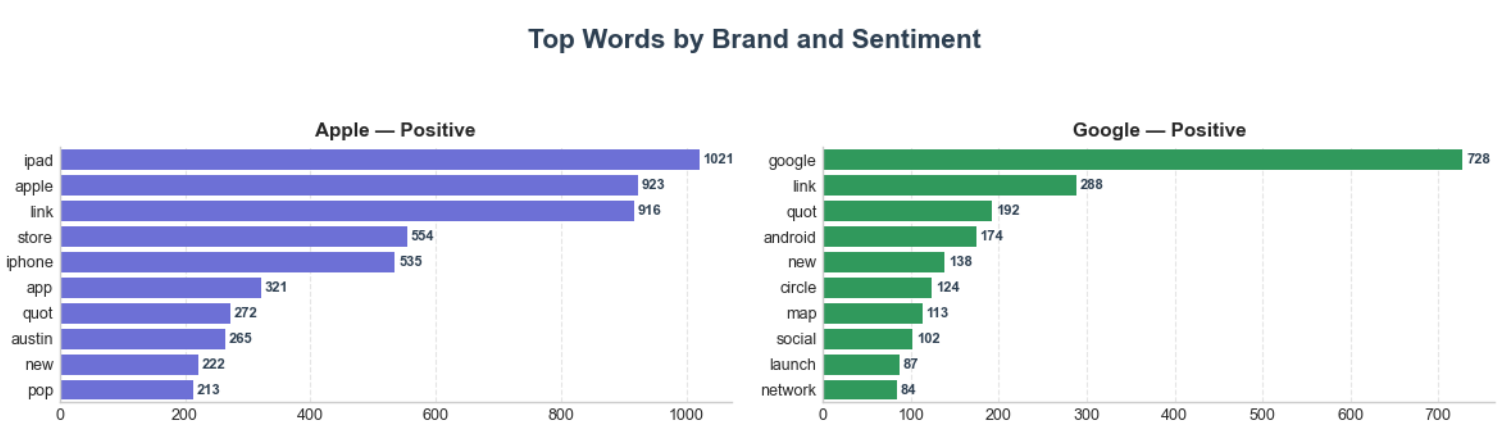
**Apple** dominates the discussion in both **volume and emotional intensity**. While neutral sentiment is most common, the significant surplus of positive over negative comments points to a high level of overall user satisfaction and strong brand affinity. **Google**, on the other hand, sees a **smaller and less volatile conversation.** Its discourse is marked by a large neutral majority, fewer positive expressions, and low negativity, resulting in a steadier, less emotional public perception.

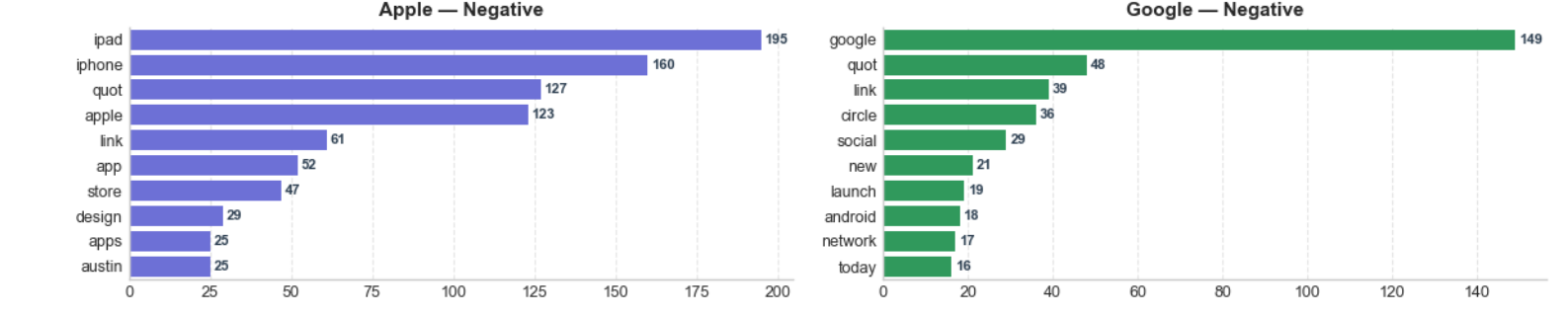
**E. Average tweet length across sentiment classes**

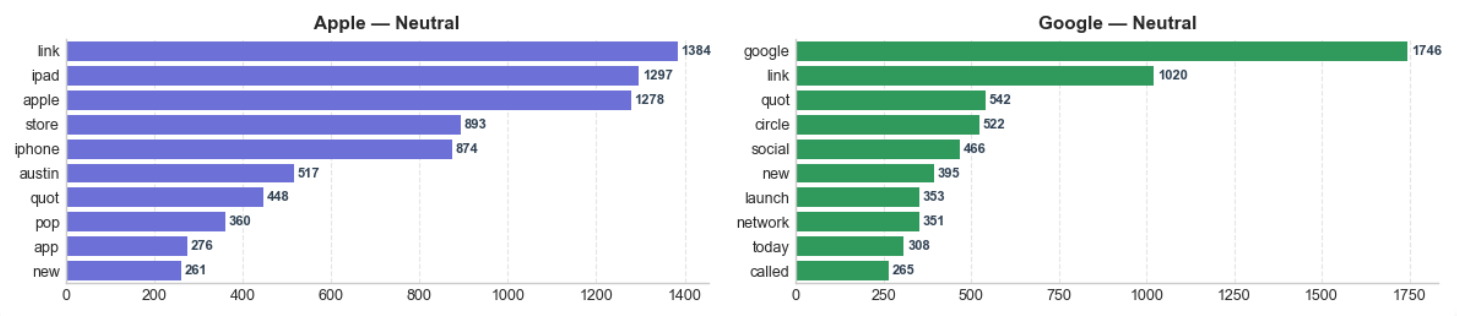
****

On average, tweets referencing **Google** are slightly longer than those mentioning **Apple**, suggesting Google discussions require more detail or context. This difference is consistent across sentiment: Apple users keep their comments **short and reactive** regardless of tone, while Google users are most **wordy and explanatory** when they are offering either praise or critique (negative tweets are the longest, followed by positive ones).

**F. Top words for each brand and sentiment**

****

****

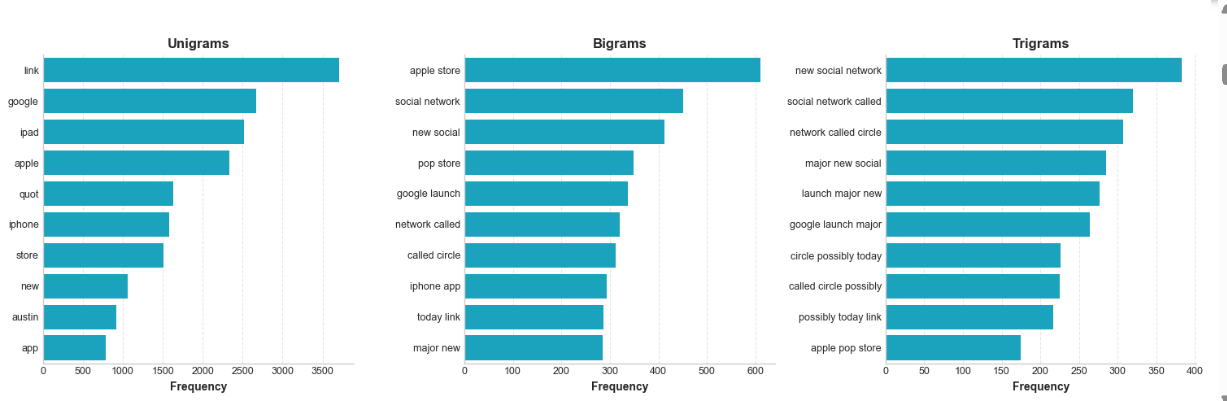
****

Discussions for both brands were heavily influenced by the **SXSW event**, suggesting that a significant portion of the conversation was **event-driven** buzz rather than routine, everyday commentary.

For **Apple**, the chatter consistently revolved around core products—the **iPad, iPhone, and Apple Store**—across all sentiment types. This indicates that user praise and criticism alike focused on these same topics, likely concerning aspects like product quality, design, or pricing.

In comparison, the **Google** conversation centered on a few key terms, including **Android, Google Maps, circle, social, and launch**. This focus reveals significant attention on Google+ and related feature rollouts. Like Apple, the reused terms in both positive and negative tweets imply that users had mixed reactions to the same subjects, whether they were offering approval or critique.

**G. Top N-Grams Analysis**

****

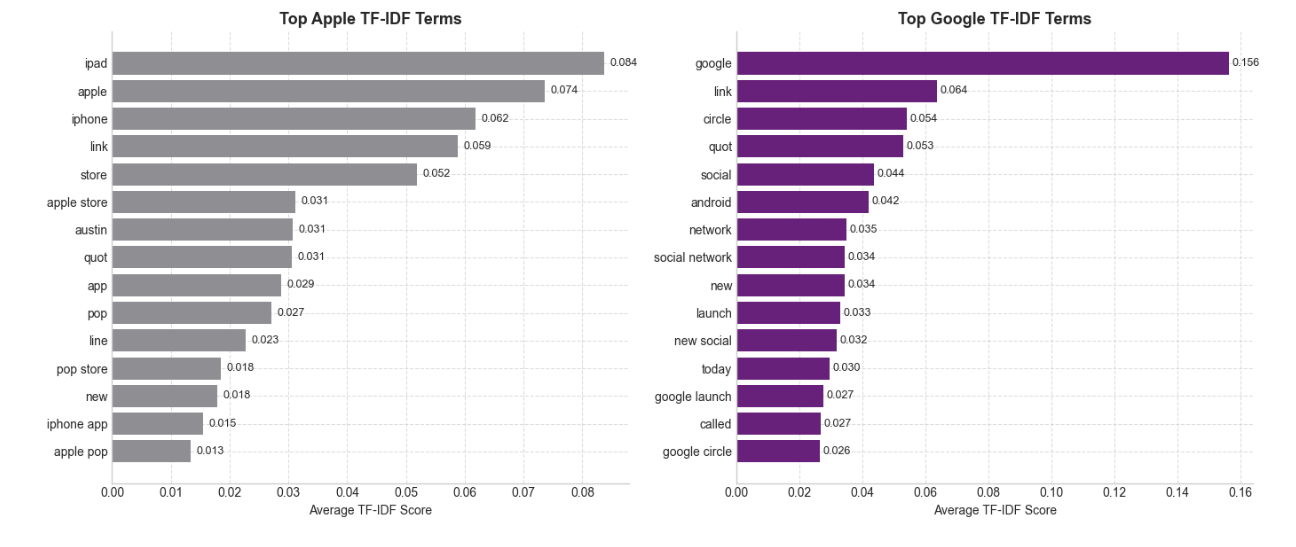
The overall conversation was significantly influenced by the **SXSW (South by Southwest) event**, driving mentions of **Apple, Google, the iPad, and new social platforms**. Linguistic analysis of phrases like 'new social network' and 'google launch major' indicates that a major **Google product announcement**, likely related to **Google+**, was trending. Separately, **Apple** maintained visibility throughout the event due to recurring **'apple store'** mentions, suggesting concurrent store-based launches or events.

| **Brand** | **Primary Focus** | **Key N-Grams** | **Conversational Style** |
| --- | --- | --- | --- |
| **Apple** | **Experiential & Product-Centric** | **'apple store', 'pop store', 'ipad', 'iphone'** | **Buzz was driven by in-person experiences, like promotional pop-up stores and product launches. The focus was tangible.** |
| **Google** | **Innovation & Platform-Driven** | **'new social network', 'network called circle', 'google launch major'** | **Discussion centered on digital features and major announcements (social networking). The focus was on future technology and community building.** |

The emotional valence of the tweets also revealed specific concerns and excitement:

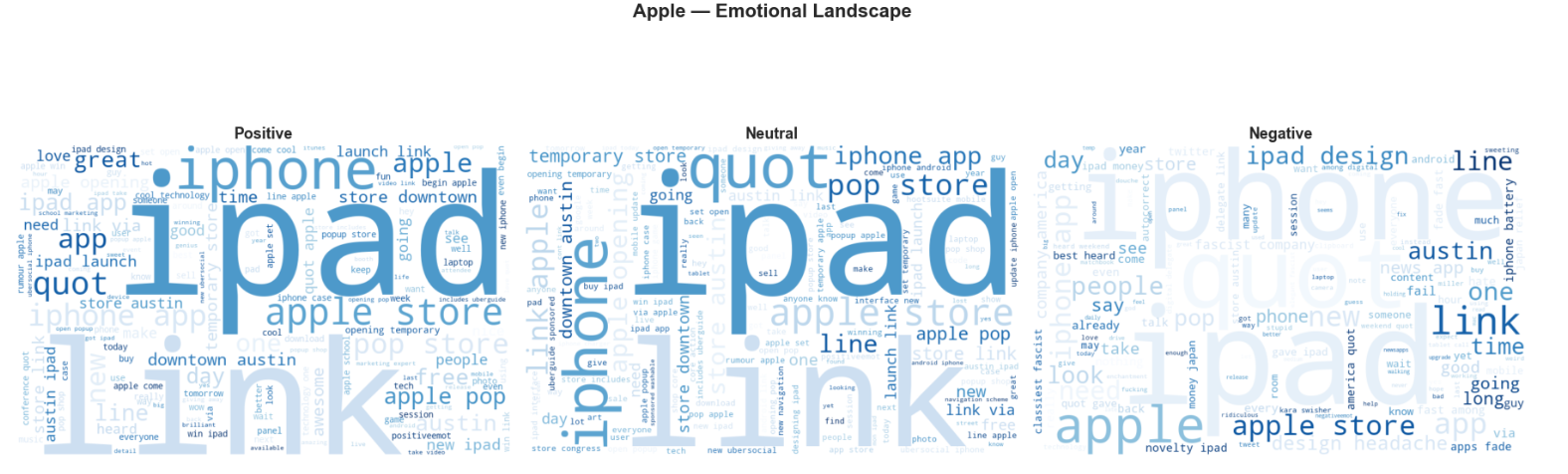
* **Negative Sentiment:** Although **sparse**, negative tweets often pointed to **design flaws, hardware frustrations** ('ipad design headache'), or serious issues like **restrictive company policies or privacy concerns** ('fascist company america'). The presence of SXSW terms in these tweets suggests the criticism was part of the ongoing, public event discourse.
* **Positive Sentiment:** This segment was characterized by **enthusiasm for Apple’s in-person events** ('apple pop store', 'sxsw apple') and **optimism regarding new technology**, including Google's ventures ('new social network') and app launches ('google map').
* **Neutral Sentiment:** Neutral posts primarily served an **informative or factual purpose**, relaying news, event updates, and links without emotional tone. This content (likely journalistic or automated) frequently included the same core SXSW and brand terms, but without the emotional charge seen in positive or negative tweets.

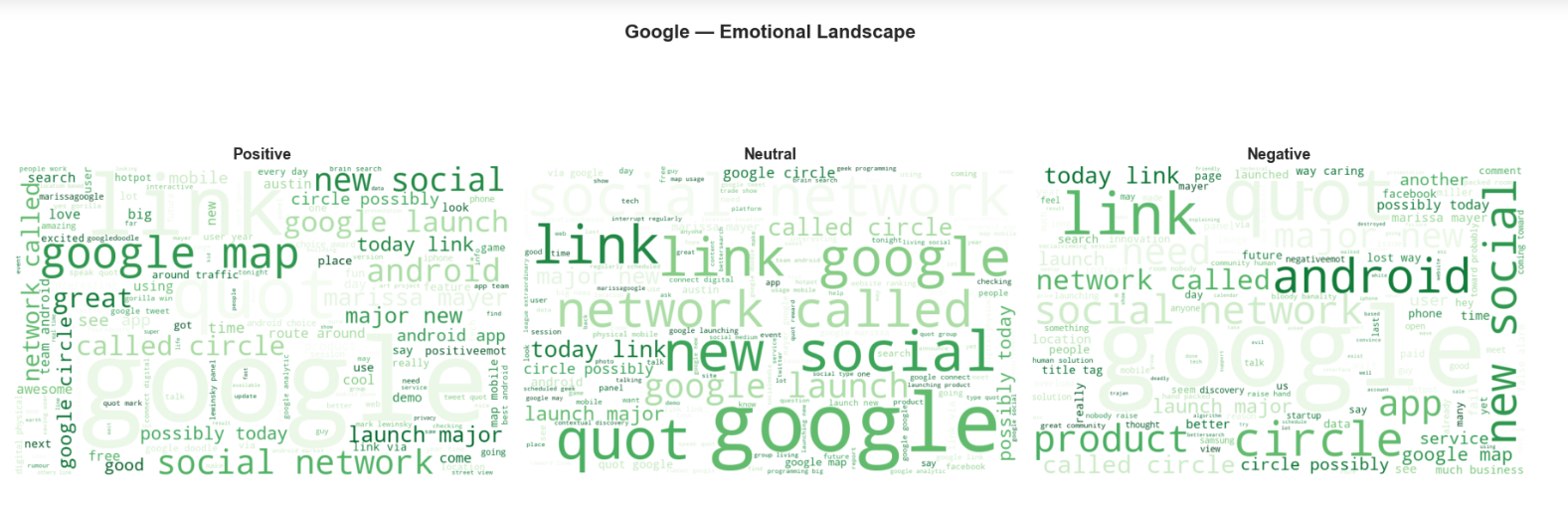
**H. TF-IDF Analysis**

****

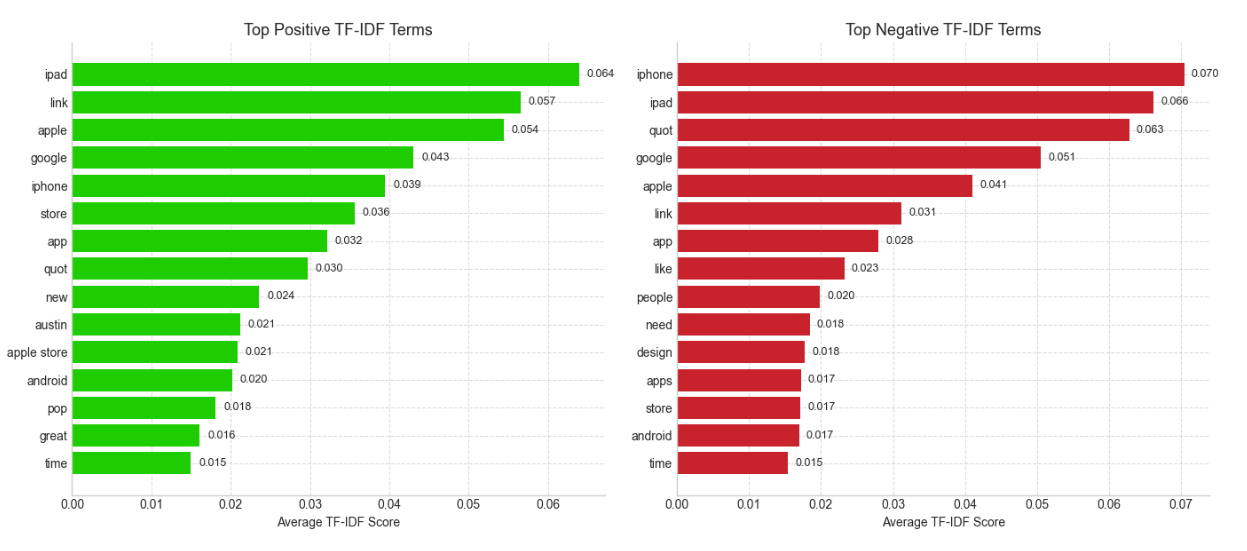
The content surrounding **Apple** is **product- and experience-driven**, with recurring keywords such as **'iPhone,' 'iPad,' and 'store'** pointing to high emotional investment in product usability and launches. Meanwhile, the **Google** discussion centers on **new technology and platforms**, using terms like **'circle,' 'social,' and 'launch'** to signal a more strategic or feature-based conversation about innovation and networks.

**I. Emotional tagging using Word Clouds**

****

****

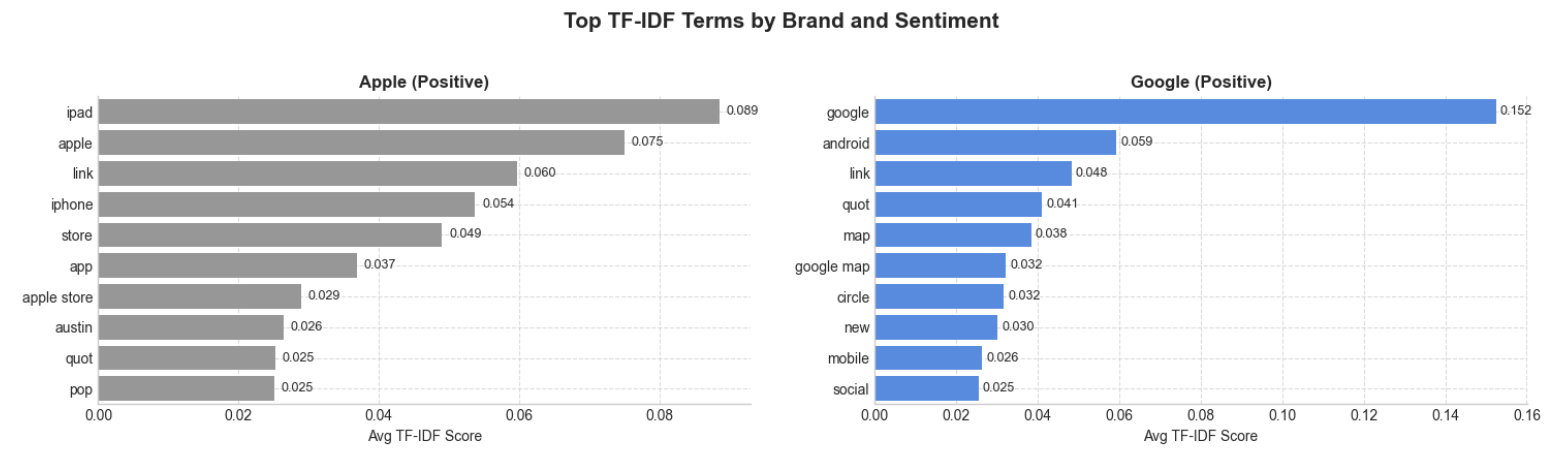
**J.** **Most common terms for positive and negative sentiments**

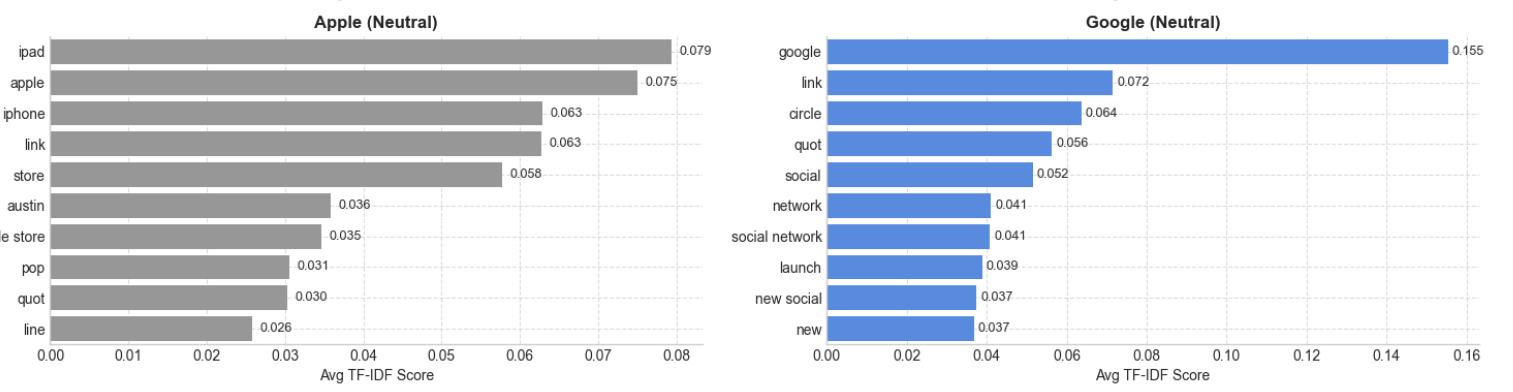
****

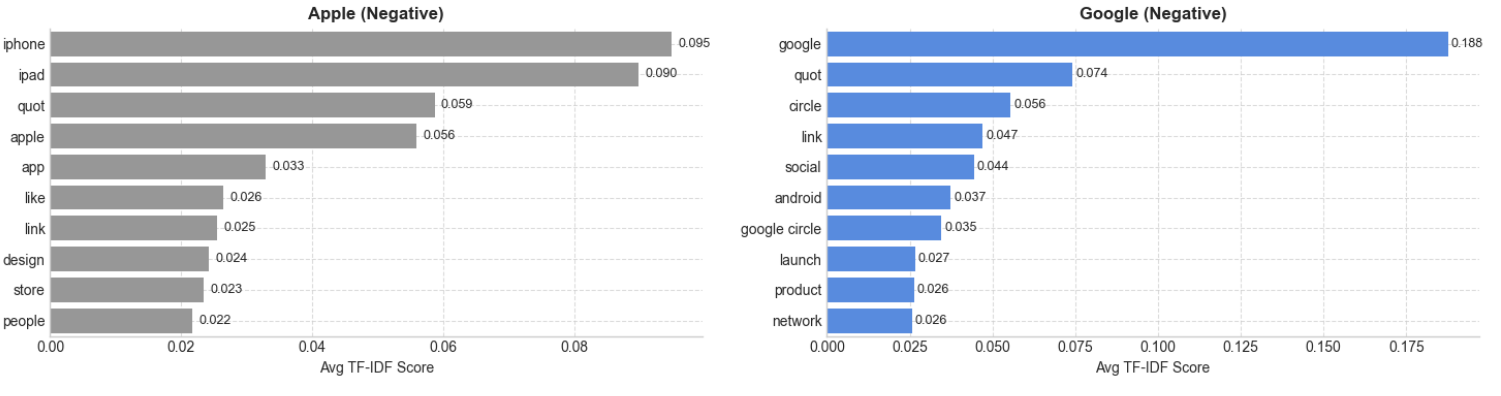
**Positive Buzz:** The key drivers of positive sentiment are the tangible **Apple products (iPhone, iPad)** and the **App Store**. Frequent use of words like **'link,' 'new,' and 'great'** demonstrates excitement tied to product announcements and feature updates. Any positive mention of Google here is typically in relation to, or admiration of, the main subjects.

**Negative Feedback:** Negative sentiment is concentrated on both brands, triggered by issues related to the **user experience**. Keywords like **'need,' 'design,' and 'apps'** point to **frustrations with performance, functionality, or restrictive design**. The high volume of **quoted criticism** (indicated by "quot" tokens) highlights that sharing complaints or existing negative press is a common form of negative engagement.

**K. Top sentiment terms associated with each brand**

****

****

****

**Apple's conversation is centered on tangible products and the user experience**, specifically the **iPhone, iPad, and Apple Store**. Users generally expressed **satisfaction with usability** but directed their **frustrations toward design or pricing**. Conversely, **Google's chatter is service-driven**, focusing on platforms like **Android, Maps, and social tools (Circles)**, where praise for **innovation** was often mixed with **complaints about reliability**.

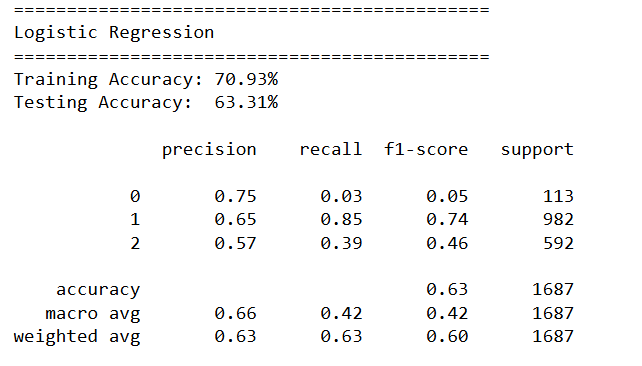
## 5. Modeling

The steps taken in this section were as follows:

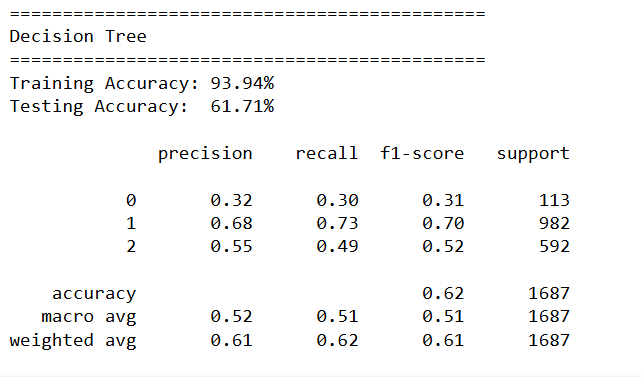
1. **Vanilla Modeling** to set up a benchmark model, using Logistic Regression, Decision Tree, Random Forest and XGBoost.
2. **Dimensionality Reduction** to apply truncated decomposition to reduce feature dimensionality.
3. **Handling Class Imbalanc**e by using SMOTE or class weighting.
4. **Hyperparameter Tunin**g using Gridsearch and Randomized Search for optimal generalization.
5. Model Evaluation and Comparison of accuracy, precision, recall, F1-score, and ROC-AUC.
6. Regularization

### 5.1 Vanilla Model Selection

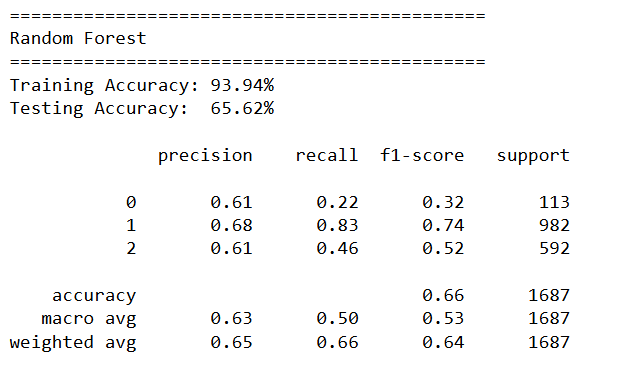
**5.1.1. Logistic Regression**



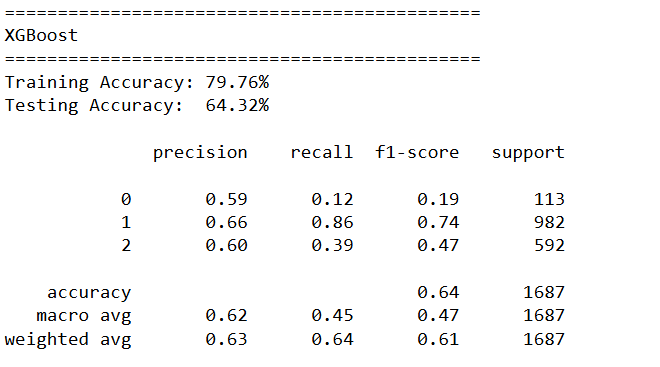
**5.1.2. Decision Tree**

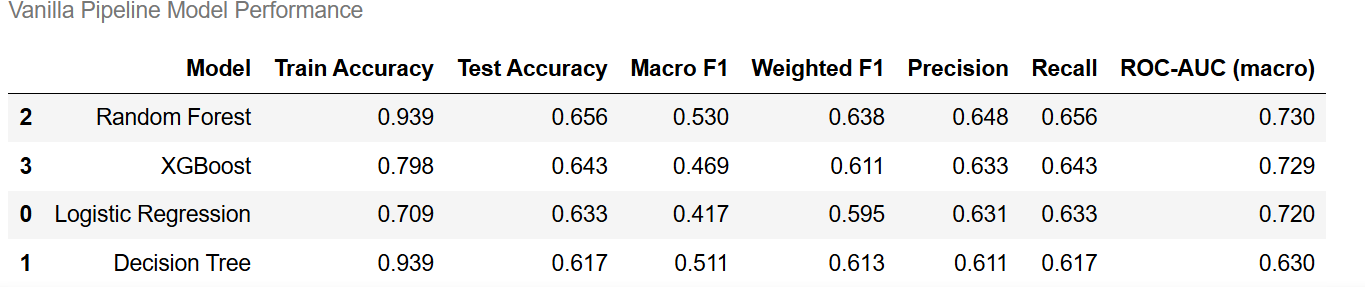


**5.1.3. Random Forest**

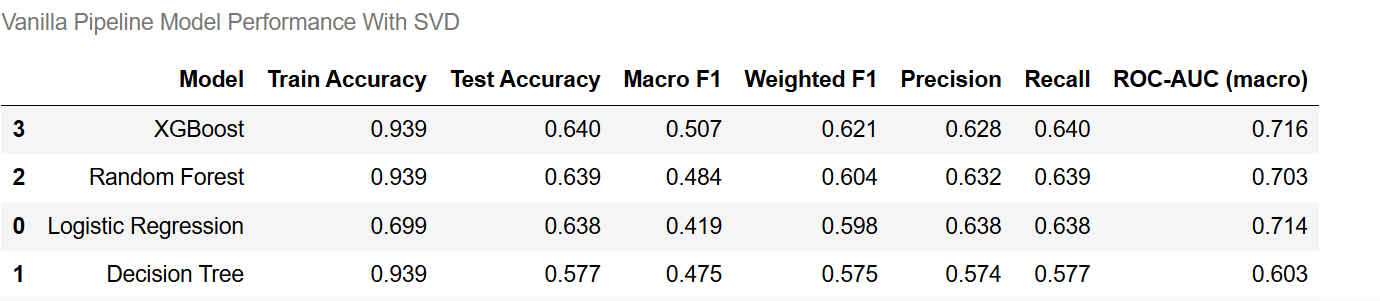


**5.1.4. XGBoost**



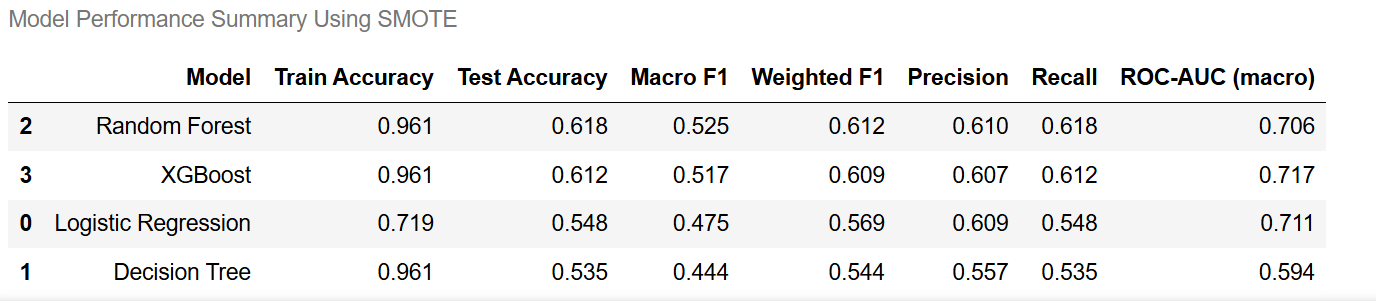


**5.2 Dimensionality Reduction**

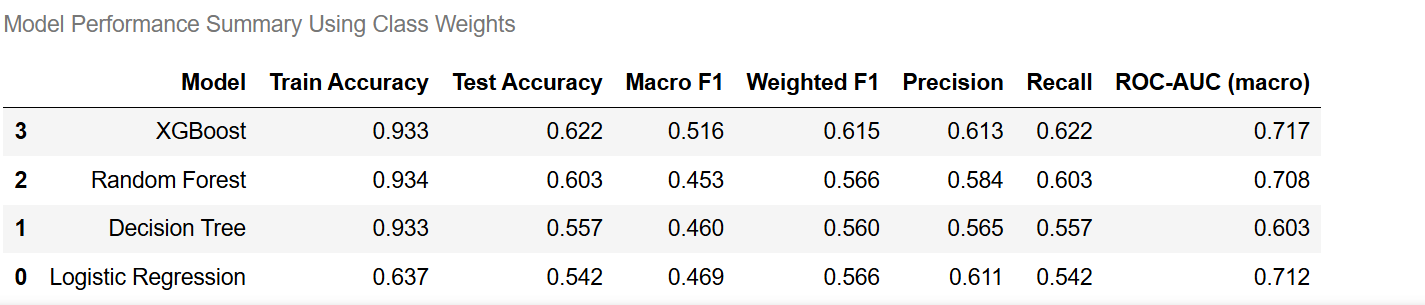
****

**5.3 Handling Class Imbalance**

**5.3.1 SMOTE**

****

**5.3.2 Class Weights**

****

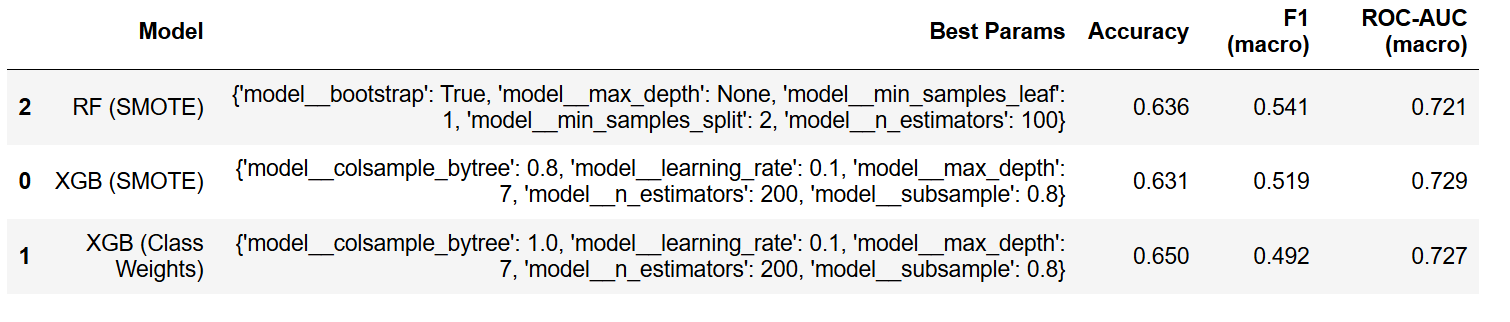
### 5.4 Hyperparameter Tuning

Initial models showed a weakness in handling **class imbalance**, resulting in predictions skewed toward the largest classes. Implementing strategies such as **SMOTE or class weights** was critical, as it successfully improved the **macro F1 scores** and ensured better overall classification reliability. Furthermore, trials with **SVD dimensionality reduction** did not demonstrate any measurable benefit to either model accuracy or the macro F1 score.

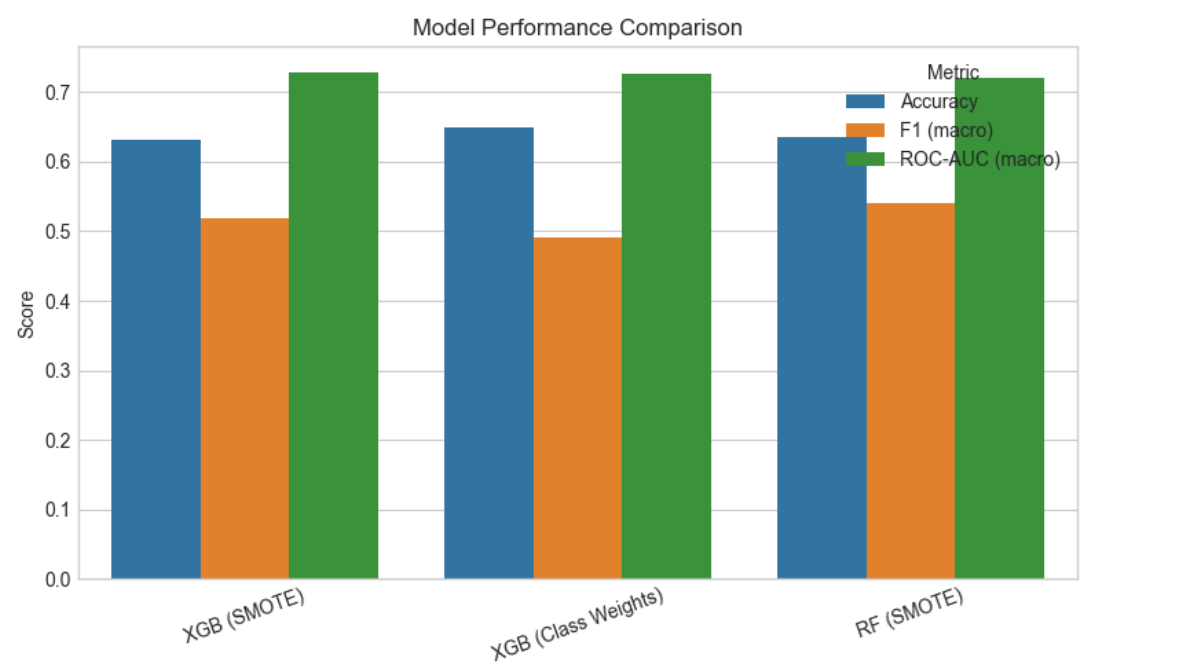
The selected models for hyperparameter tuning that provided a balance between robust **handling of class imbalance**, **high macro F1**, and **good overall predictive performance** were:

1. XGBoost(SMOTE)
2. XGBoost(Class Weights)
3. Random Forest (SMOTE)

After tuning the results were as follows



**5.5. Model Comparison**



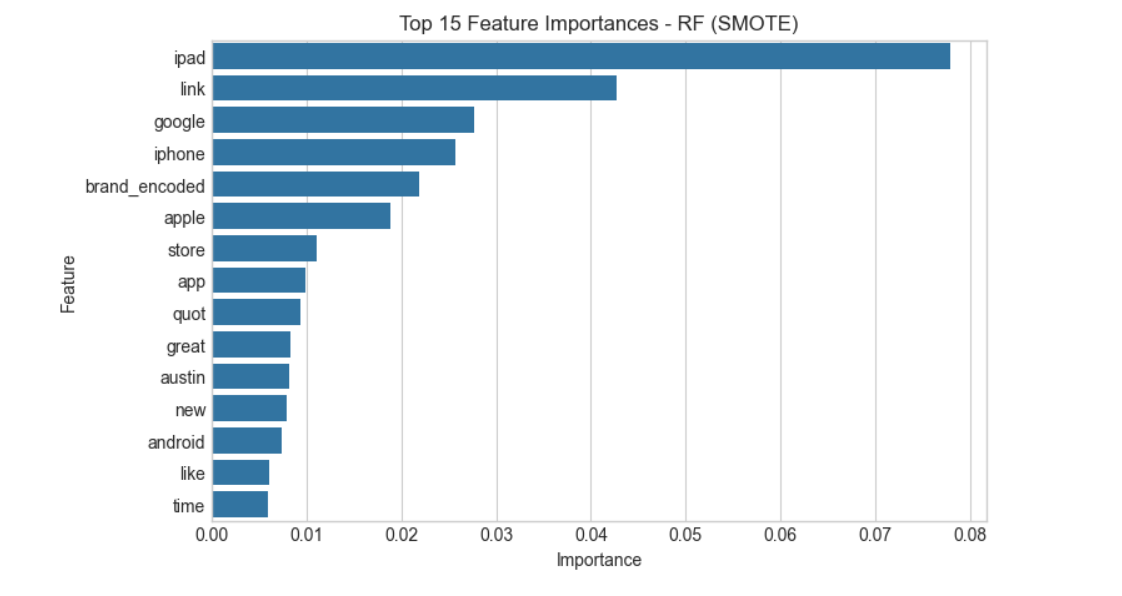
**Key Findings:**

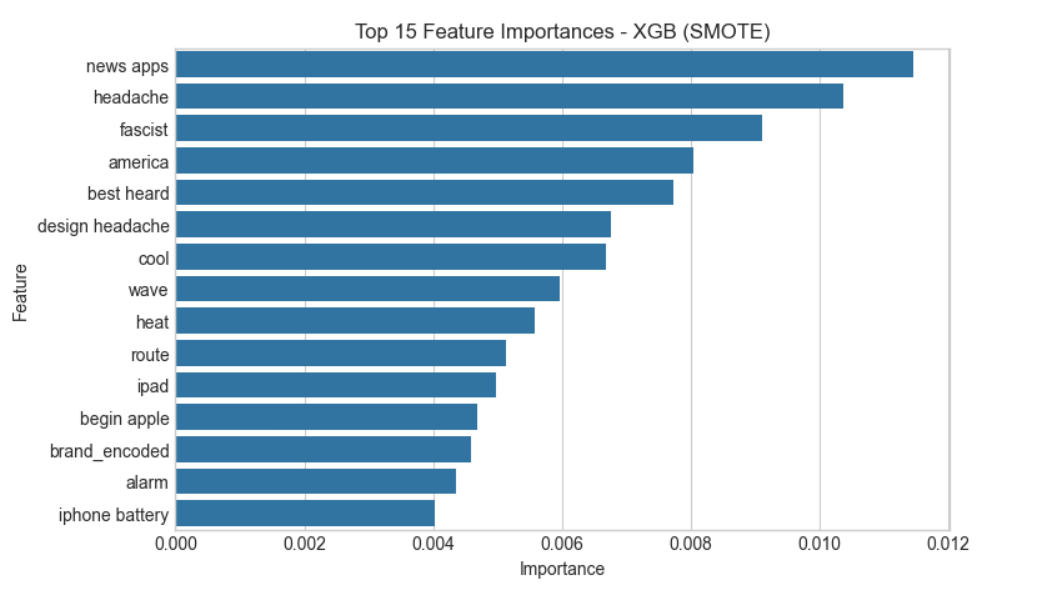
* The metric consistently performing the highest across all models is the **ROC-AUC (macro)**, which measures the models' ability to distinguish between all classes. All three configurations achieved an **ROC-AUC above 0.70 (or 70%)**, indicating strong discriminative power.
* If the priority is simply overall correct predictions (where predicting the majority class correctly is highly rewarded), **XGB (Class Weights)** is the winner due to its highest **Accuracy** (~ 65%).
* If the priority is achieving **balanced performance** across all minority and majority sentiment classes (which is critical in imbalanced datasets), the **Random Forest (SMOTE)** configuration is the clear winner, thanks to its superior **F1 (macro) score** (~54%).

**5.6. Feature Importance**

The two best models selected for feature importance after retraining were:

1. Random Forest (SMOTE)
2. XGBoost (SMOTE)





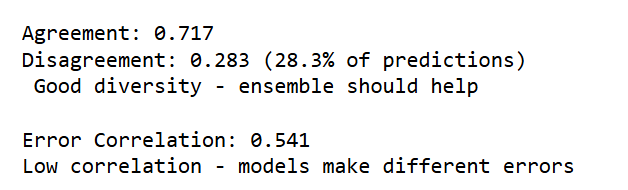
Analysis of the top-performing models, RF (SMOTE) and XGB (SMOTE), revealed two distinct feature strategies. The **Random Forest model is driven by direct product and brand keywords** like **'ipad,' 'link,' and 'iphone'**. The **XGBoost model**, however, relies on more **subtle, emotionally charged, and context-specific terms** such as **'fascist' and 'headache'**. Because each model leverages different aspects of the dataset for classification, we have strong justification to **implement ensemble techniques** to harness their combined strengths.

**5.7. Regularization**

To address overfitting the models were regularized to give the results below:



### 5.8. Model Diversity Check



The goal of this check is to see if the models are "thinking" independently. If two models make the exact same mistakes, combining them is pointless. You want them to make different mistakes so that one model can correct the error of the other.

### Agreement and Disagreement

These metrics quantify how often the predictions of the two models align on the test data:

* **Agreement (0.717 / 71.7%):** This is the percentage of time the two models produce the **same classification** (e.g., both predict 'positive' or both predict 'neutral') for a given data point. A high agreement suggests both models have captured the major, obvious patterns in the data.
* **Disagreement (0.283 / 28.3%):** This is the percentage of time the models **disagree on the classification**. This 28.3% represents the **potential value-add** of the ensemble. This large disagreement space is where a voting or stacking ensemble method can step in and make a final, more informed decision, resolving the conflict.

**Conclusion:** The notation "Good diversity" confirms that a $28.3\%$ disagreement rate is substantial enough to make ensembling worthwhile.

### 5.8. Ensemble Stacking and Voting

**5.8.1 Stacking**

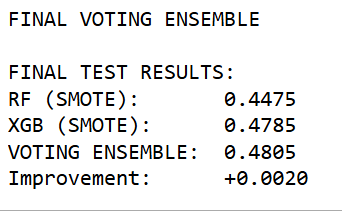
The results obtained were::

**Stacking F1: 0.4312 (+/- 0.0177)**

1. **Stacking F1 (0.4312):** This is the final **Macro F1 Score** achieved by the combined, meta-learned ensemble model. The Macro F1 Score is the most important metric for our imbalanced sentiment data, as it ensures performance is good across *all* classes (positive, negative, and neutral).
2. **Standard Deviation (0.0177):** This number represents the **uncertainty or variability** of the F1 score which is usually measured using cross-validation. A small standard deviation (like 0.0177) is good, indicating that the model's performance is **stable and reliable** regardless of which specific subset of the data it is tested on.

Since the **Stacking Ensemble's F1 score (0.43)** is **low** the ensemble is actually hurting performance instead of improving it and we therefore recommended Voting Ensemble

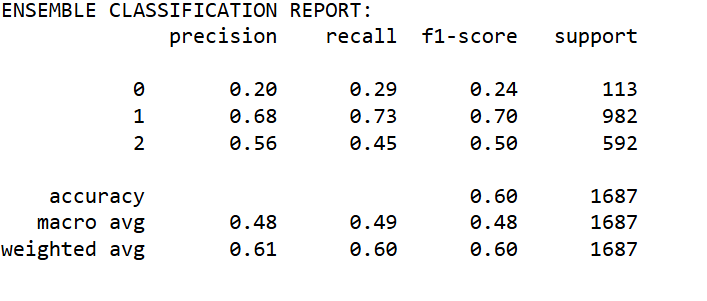
**5.8.2 Voting**

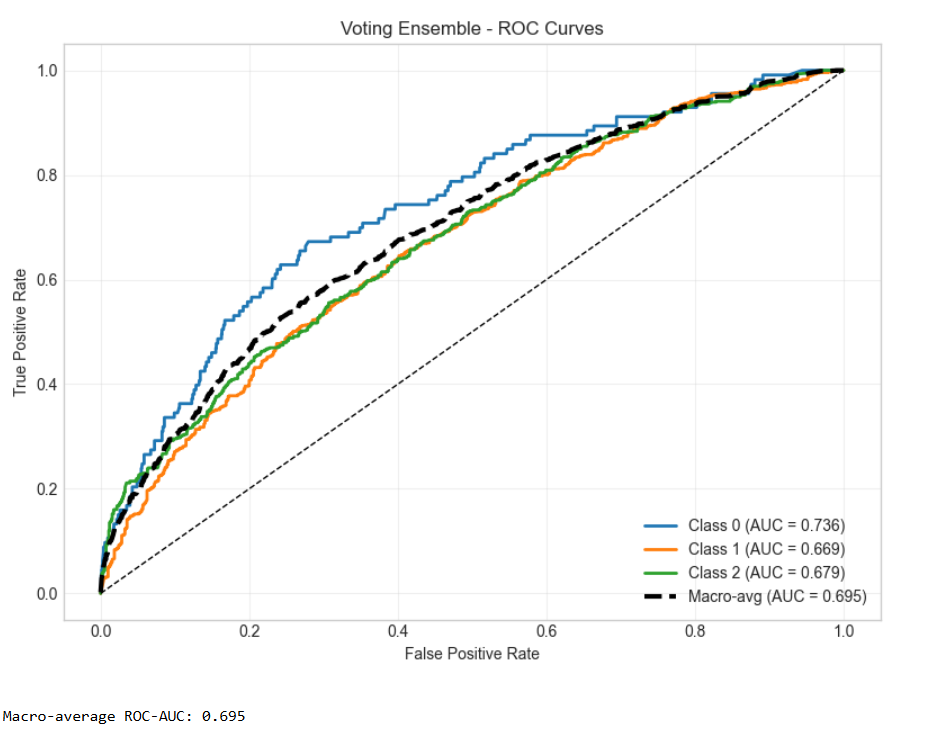
****

1. **Best Base Model: XGB (SMOTE)**
   * The **XGBoost model** (with SMOTE) is the single best-performing classifier, achieving a Macro F1 score of **0.4785**. This is your benchmark score.
2. **Ensemble Performance**
   * The **VOTING ENSEMBLE** achieved a score of **0.4805**.
   * This is a very small, but positive, **improvement of +0.0020** over the best individual model (XGBoost).

The results confirm that combining the two models through a **Voting Ensemble** was a successful strategy, though the gains were minor.

* The ensemble successfully corrected just enough errors from the individual models to **outperform the best single model**.
* This modest improvement reinforces the earlier diversity check: the models were making different mistakes, allowing the simple "majority vote" to slightly refine the final prediction.





The ensemble didn’t help much as:

1. Predictions where models differ: **478 (28.3%)**
2. On disagreed samples:

**RF correct: 125/478 (26.2%)**

**XGB correct: 265/478 (55.4%)**

It was recommended that we use **XGBoost (SMOTE**) as the final model (F1: 0.4822) as the ensemble provides no benefit in this case.

The regularization reduced overfitting but also made models more conservative and similar to each other.

## 6. Deployment????