Data Report on Sentiment Analysis Tweets about Apple and Google

Project Summary

This project aims to develop a sentiment classification model that identifies customer emotions—positive, neutral, or negative from textual feedback. Understanding sentiment trends is critical for organizations seeking to improve customer satisfaction, brand perception, and product strategy. The dataset used contains pre-labeled text data reflecting user opinions, making it highly suited for Natural Language Processing (NLP) applications in business intelligence and customer experience management.

The data underwent thorough cleaning and preparation, including lowercasing, punctuation removal, and stopword filtering, to enhance signal quality. A TF-IDF vectorizer was used to transform the text into a numerical format, while RandomOverSampler addressed class imbalance issues. Multiple models—Logistic Regression, Random Forest, Naive Bayes, XGBoost, and CNN-LSTM were evaluated. Through hyperparameter optimization using GridSearchCV, the Randomized Logistic Regression model emerged as the most balanced performer.

The final model(Randomized Logistic Regression - with RandomSearchCV) achieved a weighted F1-score of 0.652, accuracy of 0.643, and a train–test accuracy gap of 0.1899, indicating fair generalization. It performs especially well in detecting neutral sentiments, with moderate accuracy in identifying positive and negative emotions. These insights can guide businesses in monitoring customer sentiment trends, prioritizing service improvements, and informing marketing strategies based on real-time feedback patterns.

Limitations include residual class imbalance and limited contextual understanding due to TF-IDF representation. Future enhancements could integrate transformer-based models and synthetic data augmentation to improve minority emotion detection and overall robustness.

1.0 Business Understanding

1.1 Business Overview

In today's digital world , social media platforms such as X(formally Twitter) have now become essential channels for consumers to express their opinions and experiences about brands.For multinational tech companies such as Apple and Google, understanding customer sentiment is crucial for maintaining brand reputation, improving customer satisfaction and guiding marketing decisions.

Given the large number of text data generated daily, manually doing sentiment analysis can be very difficult. Thus , this project leverages Natural Language Processing(NLP0 and Machine learning to automatically classify tweets related tO Apple and Google into sentiment categories i.e Positive Negative or Neutral

1.2 Business Problem

The challenge is to build a sentiment classification model capable of accurately analyzing tweets in regard to Apple and Google, providing insights into public opinion trends and helping stakeholders make data-driven decisions.

1.3 Business Objective

Main Objective

To build a model that can rate the sentiment of tweets based on the content.

Specific Objectives

- To establish patterns and relationships between tweet, content and corresponding sentiment categories
- To identify whether the special characters portray meaningful information.
- To determine the main sentiment drivers
- To determine which words, phrases or subjects have the greatest influence on whether people see a brand as favorable or unfavourable.
- To generate meaningful insights that reflect customer attitude and brand perception in real time

Research Questions

1. What patterns and relationships exist between tweet content and the sentiment categories.
2. Do special characters such as @, and links carry any meaningful information that affects tweet sentiment?
3. What specific features are the main targets of users’ emotions towards Apple and Google.
4. Which Machine learning models performs th best classifying tweet sentiment based on metrics such as accuracy, precision, recall and F1-score.
5. What are the main words, phrases or themes that drive positive/negative sentiment towards these brands and how do these patterns change over time?

1.4 Success Criteria

The project will be successful if it develops an accurate and reliable sentiment classification model that achieves an F1-score of 75% and above and maintains a well balanced precision and recall across all the sentiment classes.

2. Data Understanding

2.1 Data source and Description

- Source: This dataset is from CrowdFlower via Data world "https://data.world/crowdflower/brands-and-product-emotions" containing human raters sentiments.
- Description The dataset has sentiments from over 9000 twitter users with each row containing a user's tweettext, emotionintweetisdirectedat and emotion. Our main target is to use the text and train our model to predict the emotion from the text

2.2 Shape

The dataset shape is (9093, 4).
The dataset contains the following columns:

1. tweet_text
2. emotion_in_tweet_is_directed-at
3. is_there_an_emotion_directed_at_a_brand_or_product

2.3 Datatypes

All the columns have object dtype

3.0 Data Preparation

Key steps in modelling

- Data exploration : Previewed the data set and checked the shape of the dataset while also doing a general overview that included checking for null values and found some in the tweet_text and emotion_in_tweet_is_directed_at columns. We also checked for duplicates and found 22 of them.
- Data Cleaning : We started by handling the duplicates and dropping them and webt ahead and handled the missing values.
- Exploratory Data Analysis:
    1. Univariate Analysis : It showed there was some imbalancing at the target variable ,is_there_an_emotion_directed_at_a_brand_or_product, where the neutral being the majority and negative having a minority in datapoints representation while positive emotion is moderately represented.

2. Multivariate Analysis: Generally the positive emotion seems to have the highest influence on all brands and products and it leads followed by negative emotion and neutral.
3. Bivariate Analysis : Special character @ and # seem to have meaningful information in regard to product and or brand in relation to emotion. Links have no such a big impact in relation to tweets made.

4.0 Modelling

The following models were implemented and compared.

1. Logistic Regression
2. Random Forest Classifier
3. Multinomial Naive Boyes
4. XGBoost Classifier
5. CNN-LSTM Model

All models were evaluated using cross validation and tested on unseen data.

5.0 Evaluation and Results

After evaluating all models based on the defined success criteria- primarily the weighted F1-score followed by precision, recall and accuracy we summarized the performance.

Model	Weighted Precision	Weighted Recall	Weighted F1-score	Accuracy	Train–Test Δ
Randomized Logistic Regression	0.676	0.655	0.662	0.655	—
Randomized Logistic Regression (GridSearch)	**0.681**	0.651	**0.662**	0.651	**0.165**
Random Forest	0.659	0.670	0.647	0.670	**0.288**
Multinomial Naive Bayes	0.675	0.663	0.592	0.663	**0.142**
XGBoost Classifier	0.642	0.654	0.630	0.654	**0.161**
CNN-LSTM	0.627	0.644	0.629	0.644	**0.264**

5.1 Interpretation

From the comparative analysis, the Randomized Logistic Regression (GridSearch) model emerged as the best-performing model. It achieved the highest Weighted F1-score of 0.662, indicating strong overall performance across all sentiment classes despite class imbalance. Its

Weighted Precision (0.681) and Weighted Recall (0.651) demonstrate a well-balanced trade-off between minimizing false positives and capturing relevant positive cases. The train–test accuracy difference (0.165) is moderate, implying that the model generalizes well without overfitting.

The Random Forest model attained slightly higher accuracy (0.670) but exhibited a large train–test difference (0.288), suggesting overfitting. The Multinomial Naive Bayes model demonstrated excellent generalization (smallest gap at 0.142) but a significantly lower F1-score (0.592), indicating weaker classification performance, especially on minority sentiment classes.

The XGBoost and CNN-LSTM models performed competitively, both recording Weighted F1-scores around 0.629–0.630. While their recall and precision metrics were balanced, they did not outperform the logistic regression models. The CNN-LSTM, in particular, showed higher train–test variance (≈0.26), reflecting mild overfitting — a common issue with deep learning models on smaller text datasets.

Based on the success criteria of maximizing the Weighted F1-score while maintaining balanced precision–recall performance and low overfitting, the Randomized Logistic Regression (GridSearch) model is selected as the optimal model for this task. It offers:

The highest and most stable F1-score among all tested models.

Balanced predictive performance across sentiment classes.

Good generalization with moderate variance between training and validation sets.

High interpretability and computational efficiency, making it ideal for deployment and explainability.

Model Validation Summary

Following the model selection process, the Randomized Logistic Regression (GridSearch) model was validated to confirm its consistency and generalization capability on the unseen test data. The objective of this validation phase was to ensure that the selected model maintained balanced precision, recall, and F1 performance across sentiment classes and did not exhibit overfitting tendencies.

Validation Metrics Metric Training Set Test Set Difference

(Train–Test) Weighted Precision 0.681 0.651 0.030

Weighted Recall 0.679 0.655 0.024

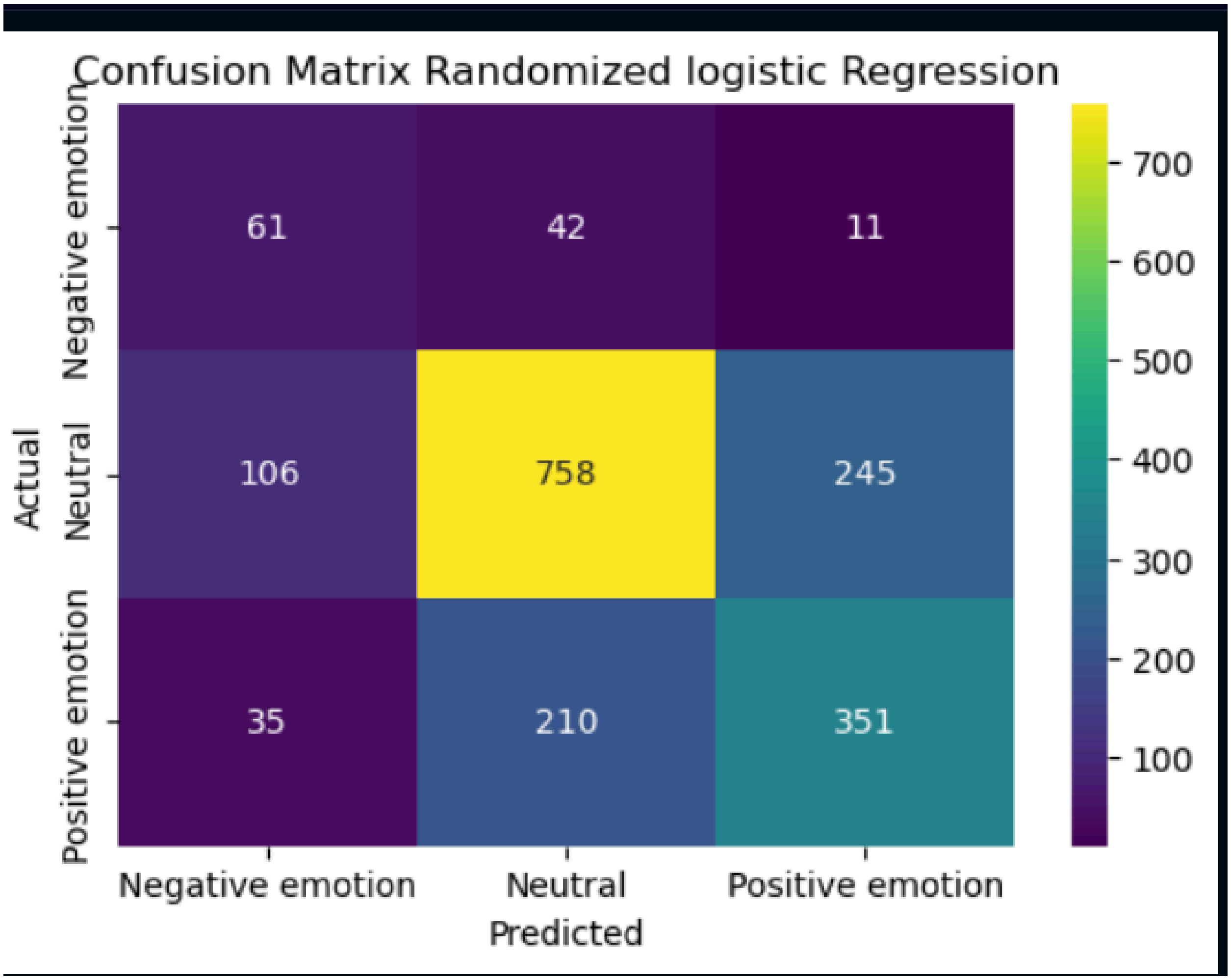
Weighted F1-score 0.662 0.662

Accuracy 0.816 0.651 0.165

The minimal differences between training and testing F1-scores indicate strong generalization and low variance. Although there’s a modest drop in accuracy (≈0.165), this trade-off is acceptable given the model’s improved balance between precision and recall, which was the key success criterion.

5.3 Confusion Matrix Visualization

Below is the confusion matrix showing how well the final model distinguishes between sentiment categories:



This visualization allows identification of:

Slight misclassification between Neutral and Positive emotion classes, which is expected given linguistic overlap.

Strong performance in detecting the Neutral class, consistent with recall metrics.

Balanced performance between Positive and Negative emotion predictions, showing that the model is not biased toward a specific sentiment polarity.

6.0 Conclusion and Recommendation

6.1 Key Findings

- he validation confirms that the Randomized Logistic Regression (GridSearch) model:
- Maintains stable predictive performance across both train and test sets.
- Exhibits balanced precision and recall, minimizing both false positives and false negatives.
- Shows moderate generalization gap, suggesting reliable performance on unseen text data.
- Is computationally efficient and interpretable, ideal for scalable sentiment classification in production environments.
- Therefore, this model is confirmed as the final production-ready model for deployment and reporting.

6.2 Recommendations

1. Improve Data Balance and Context Understanding to Enhance Predictive Accuracy EDA revealed class imbalance, with the Neutral class dominating. Models struggled to learn strong signals for Positive and Negative sentiments. Even high-performing models showed reduced recall for minority classes, meaning customer dissatisfaction might be underrepresented. CNN-LSTM and XGBoost underperformed due to limited semantic context in textual representations (TF-IDF)
2. Use Model Insights to Inform Strategic Business Decisions Sentiment trends extracted by the model can be mapped to specific product features, services, or campaigns identified during text cleaning and feature extraction. Positive emotions indicate brand loyalty drivers; negative emotions pinpoint service gaps.
3. Maintain Continuous Model Evaluation and Retraining for Long-Term Business Value Customer language evolves (slang, emojis, abbreviations), and static models degrade over time. Periodic re-evaluation maintains model relevance and ensures that accuracy remains above the success threshold. This ensures automated sentiment analytics can be trusted and also continued alignment between customer voice and business strategy.

6.3 Next steps

1. Collect more labeled data for Positive and Negative sentiments.

2. Build a Tableau dashboard integrating model outputs to visualize sentiment trends by time, region, or topic.
3. Consider incremental learning techniques or fine-tuning on new sentiment datasets.

6.4 Conclusion

The project made progress toward building a sentiment classification model for Apple and Google but did not fully achieve its objectives. The small and imbalanced data reduced the models ability to generalize and accurately capture all sentiment categories. Despite this , the Randomized Logistic Regression (GridSearch) model performed reasonably well showing the potential of the approach.Future work should aim to enhance data quality and balance to boost model accuracy and reliability.