TWITTER SENTIMENT ANALYSIS

BY
MARGARET MITEY

OVERVIEW

- In our tech-driven world, Twitter is a hub for opinions and sentiments.
- Tech advancements and natural language processing (NLP) help us analyze sentiments at scale.
- Twitter's 368M monthly users make manual analysis impractical.
- Machine learning (ML) models help gain insights into public sentiment on brands.
- Apple and Google products are some of the most sought after and used technology applications. Apple are innovators in hardware and software devices. and Google offer services such cloud computing and artificial intelligence (AI).
- The project aims to build a model that can rate sentiment of tweets based on its content

Problem Statement

In today's digital landscape, gauging customer sentiment is a challenge for Apple and Google. The existing systems lack real-time content-based ratings from users, hindering their ability to understand customer satisfaction and categorize opinions effectively.

Chemami Ent. specializes in analyzing customer feedback across brands.Our system provides real-time insights into customer sentiment. Brands can adapt products to meet customer needs effectively.

Objective

To build a model that can rate sentiments of a tweet based on it’s content of Apple and Google products.

Data

The dataset was sourced from CrowdFlower via https://data.world/crowdflower/brands-and-product-emotionsis from the year 2013 and has 8,721 tweets.

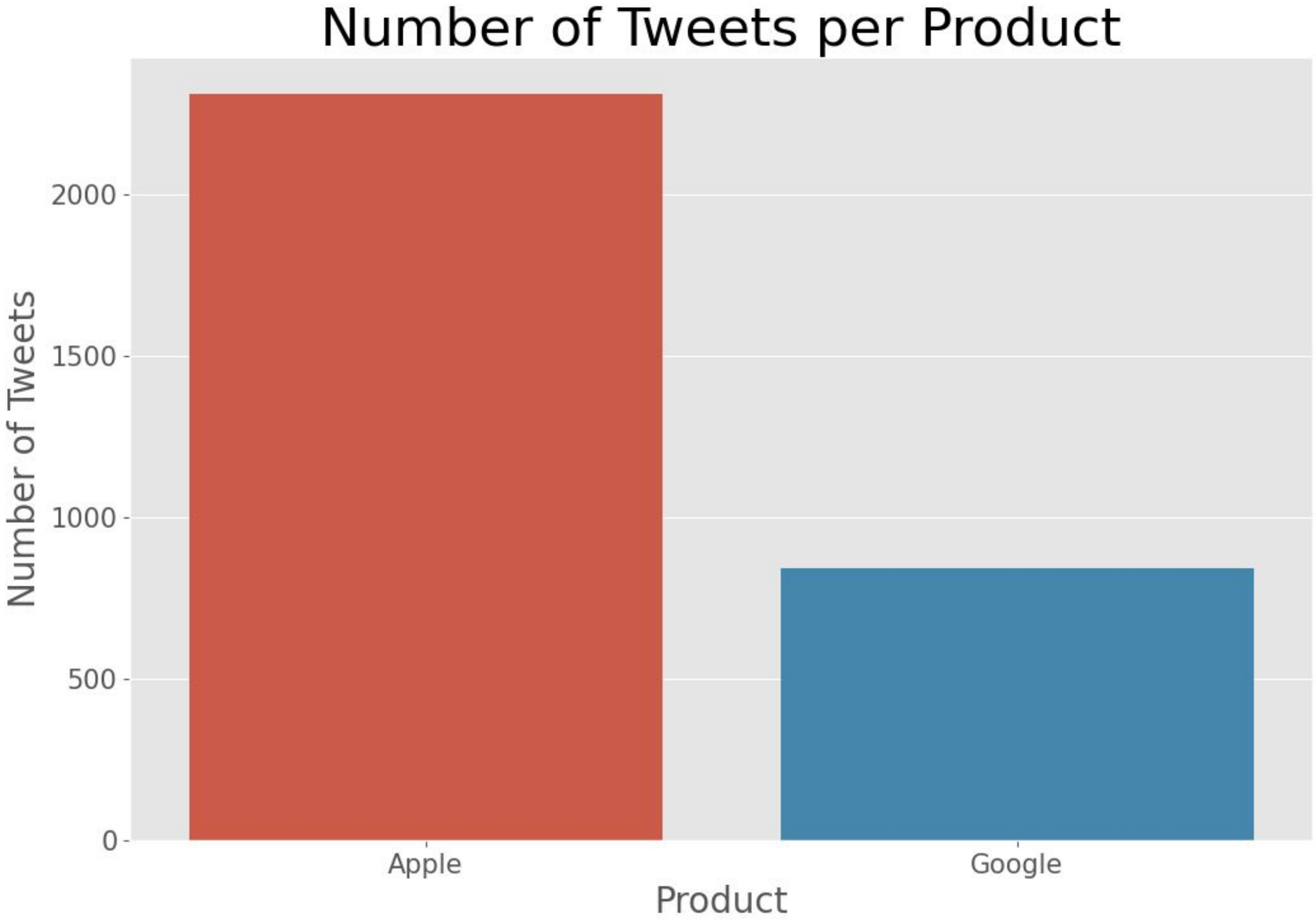
Data Preparation and Presentation

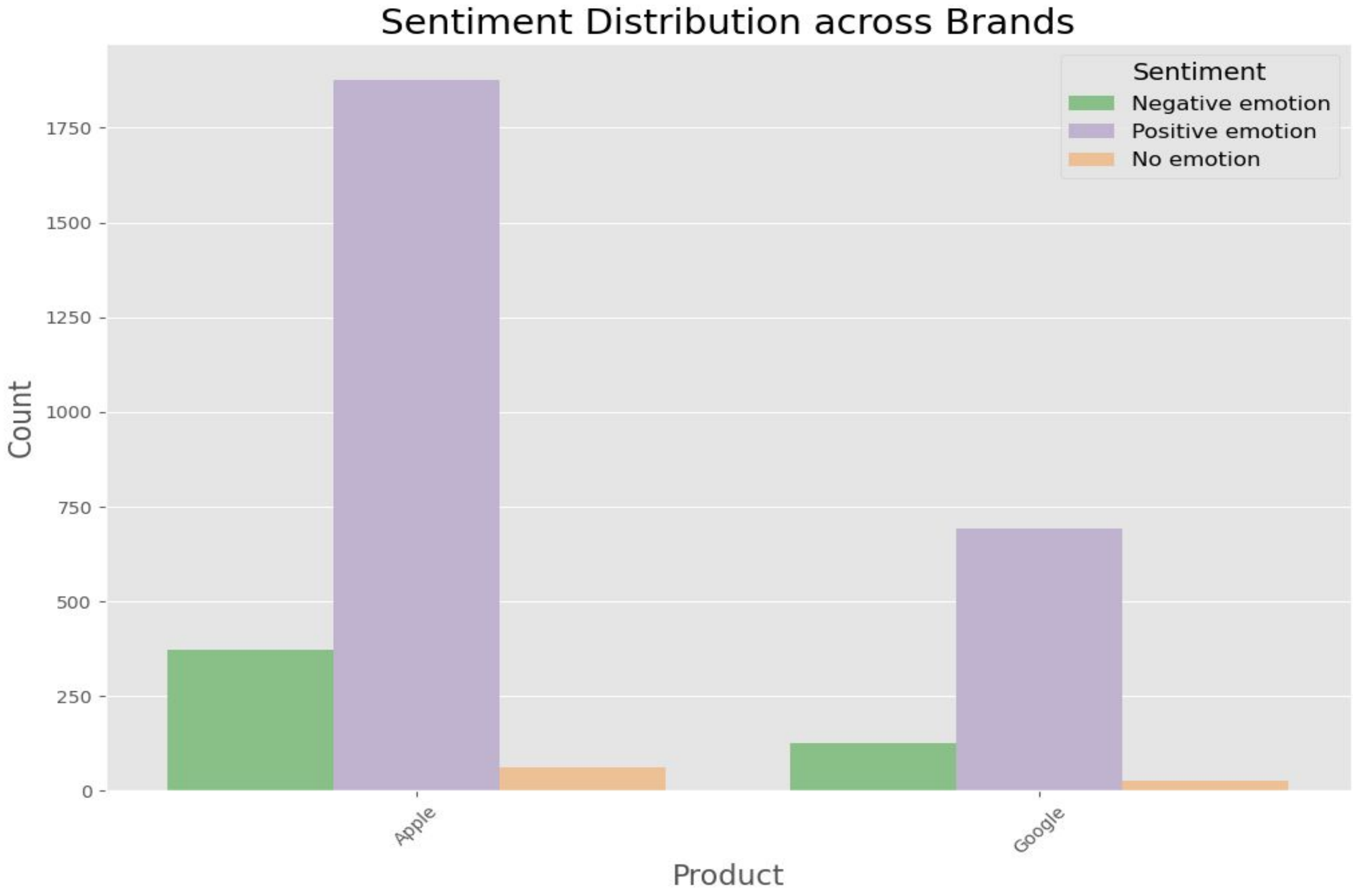
To make it easy to read and analysis the data, the column were renamed and the sentiment column was simplified to have no emotion, negative emotion and positive emotion.

Missing values and duplicates were removed from the dataset.

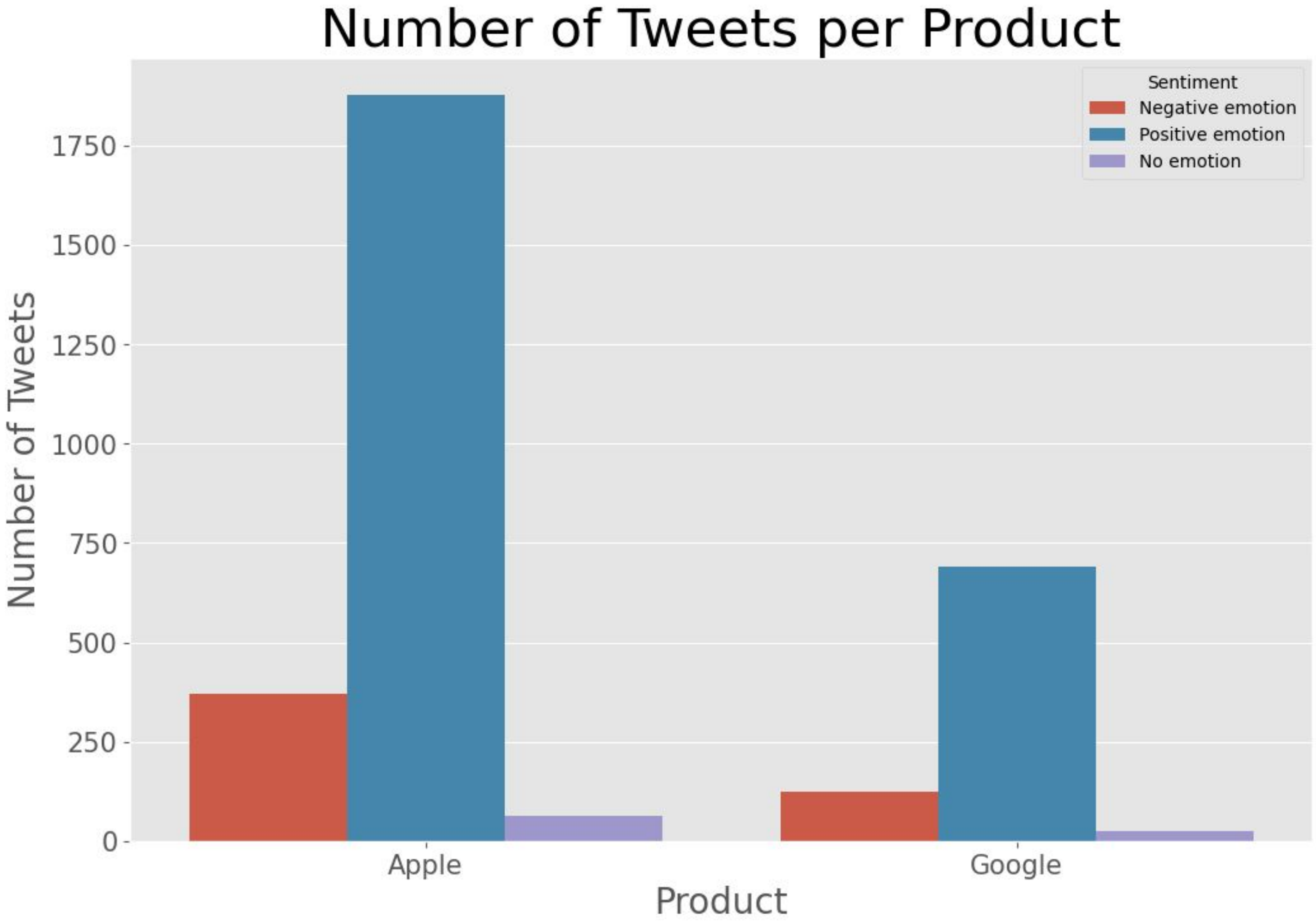
Cleaning and preprocessing the data so as to the product column are correctly labeled.

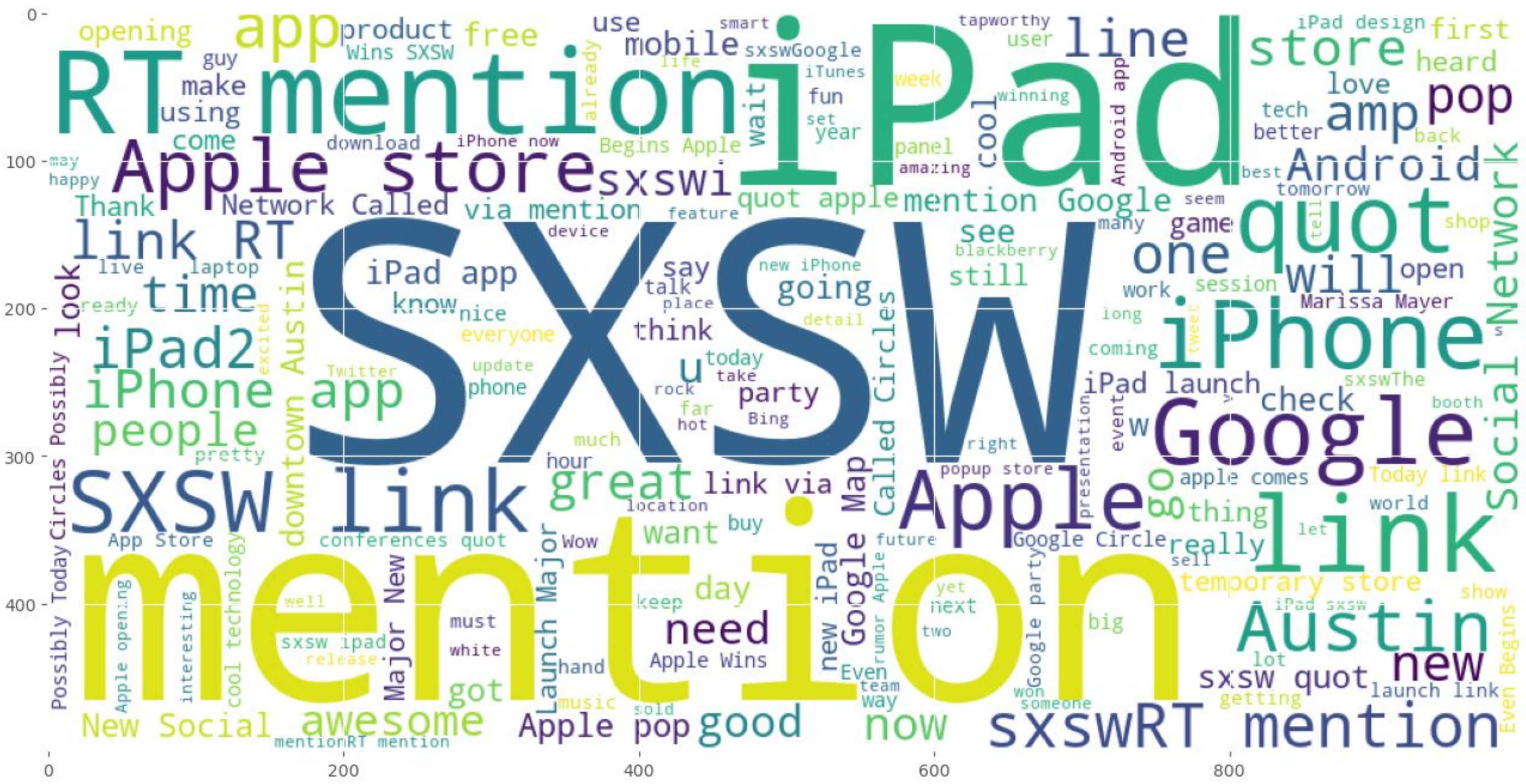
Exploratory Data Analysis is vital to gain insight, detect patterns and understanding the underlying structure of the dataset.


Number of Tweets per Product
Number of Tweets
2000
1500
1000
500
0
Apple
Google
Product



Sentiment Distribution across Brands

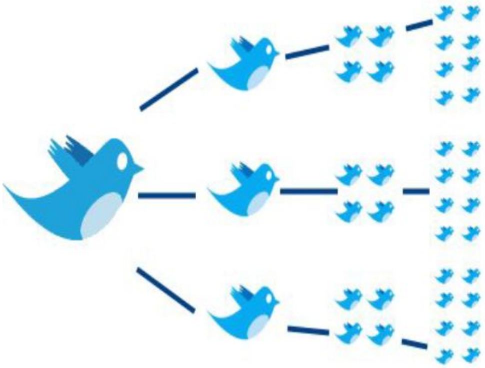

Number of Tweets per Product
Sentiment
Negative emotion
Positive emotion
No emotion
Number of Tweets
0
250
500
750
1000
1250
1500
1750
Apple
Google
Product



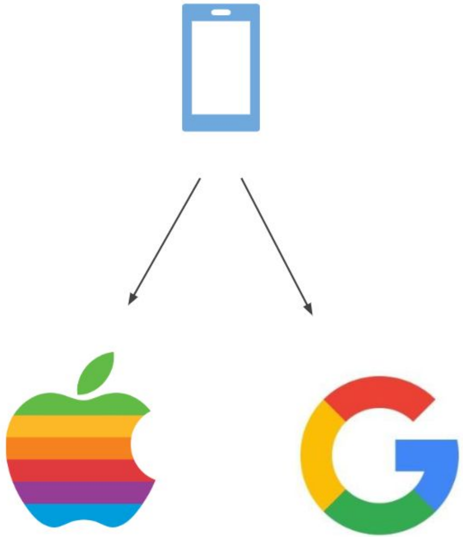
Word Cloud is depicting the frequent words presented as the largest and the less common words as the ones that are small.

Dataset

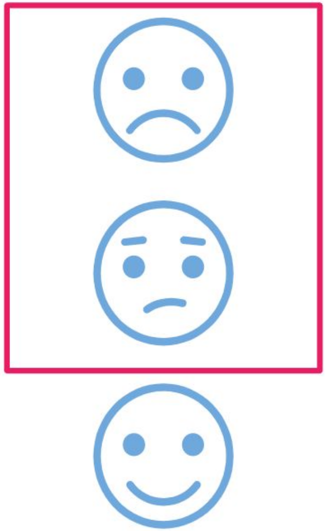
Tweet



Product



Sentiment



Data Preprocessing

- This is a very crucial part of data preparation to prepare for modeling.
- This involves removing of hashtags, retweets, hyperlinks, punctuations, white spaces and non-letter from the text.
- The next step is tokenization, stemming and lemmatization to convert it to a format that can be applied for modeling..

Modeling

Prior to developing and employing models. The data has to undergo a vectorization process of converting text format into numericals format.

For this project, the models used were suitable for analysis of datasets that were in text format. The models are:

1. Multinomial Naive Bayes Model
2. Random Forest Model
3. Support Vector Machine (SVM), with hyperparameter tuning

Evaluation of the Models

1. Multinomial Naive Bayes

   It serves as a good baseline model for sentiment analysis. The text data has been converted into tokens.

   In the output below, the accuracy of the model is 84%, which means it correctly predicted the sentiment column for 84% of the sample in the dataset

```
Accuracy: 0.8431061806656102
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        13
           1       0.86      0.18      0.30       101
           2       0.84      0.99      0.91       517

    accuracy                           0.84       631
   macro avg       0.57      0.39      0.40       631
weighted avg       0.83      0.84      0.79       631
```

Multinomial Naive Bayes

```
Accuracy: 0.8541996830427893
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        13
           1       0.81      0.29      0.42       101
           2       0.86      0.99      0.92       517

    accuracy                           0.85       631
   macro avg       0.55      0.42      0.45       631
weighted avg       0.83      0.85      0.82       631
```

Random Forest Model
This is a step-up from the baseline model to check for performance improvement.

```
Fitting 3 folds for each of 27 candidates, totalling 81 fits
Best Hyperparameters: {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
Final Model Accuracy: 0.8716323296354992
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.08      0.14        13
           1       0.79      0.42      0.55       101
           2       0.88      0.98      0.93       517

    accuracy                           0.87       631
   macro avg       0.89      0.49      0.54       631
weighted avg       0.87      0.87      0.85       631

Confusion Matrix:
 [[  1   1  11]
 [  0  42  59]
 [  0  10 507]]
```

Support Vector Machine with hyperparameter tuning.
The accuracy score is 87% showing improvement in performance from the baseline model.

Conclusion

The aim of this project was to build a model that could rate the sentiment of tweets based on its content for Apple and Google products . In order to do so, several multiple classification models were tested and identified. Support Vector Machine with hyperparameter tuning had the highest performance with an accuracy of 87% and f1-score of 93%. It performs well in identifying tweets with positive emotion.

The model will allow stakeholders to identify priority users to target (negative and no emotion) feedback and target their ads accordingly.

Recommendation

Stakeholder should focus their efforts on engaging with Twitter users who have expressed negative sentiment towards Apple and Google products as to address any concerns or negative experiences these users may have had so as to retain them.

Additionally, those with neutral views toward Apple and Google products should not be overlooked, as they represent a potential customer base that can be further cultivated and present an opportunity for growth, thus expanding the customer base.

The machine learning models employed in our sentiment analysis can benefit from ongoing refinement and optimization.

THANK YOU