

Twitter Sentiment Analysis

By Group 1:
Grace Wangui
Salma Mwende
Stephen Muema

Business understanding

Social media is a major channel for customers to share opinions and emotions. Platforms such as Twitter generate a large volume of unstructured text data. Monitoring sentiments helps technology companies like Apple and Google Improve products, enhance customer satisfaction and manage brand perception.

Business Problems

Negative sentiment about products or services can quickly harm brand perception, reduce customer loyalty, and impact sales. This project aims to address this problem by using Natural Language Processing (NLP) and machine learning techniques to automatically classify and analyze the sentiment expressed in tweets about Apple and Google products.

Key Questions

- How can a machine learning model classify tweets as positive or negative?
- How to clean and preprocess raw Twitter text data?
- How to explore and visualize sentiment distribution effectively?
- Which evaluation metrics ensure reliable model performance?

Sentiment Distribution

(Before Simplification)

Sentiment	Count	%
No emotion	5,389	59%
Positive	2,978	33%
Negative	570	6%
I can't tell	156	2%

(After Mapping)

Sentiment	Count	%
Neutral	5,531	61
Positive	2,97	3.275
Negative	569	627

The tables show the sentiment distribution, highlighting class imbalance with most tweets being neutral and far fewer negative or uncertain.

Data Preparation

To prepare the Twitter dataset for analysis, we first clean the tweets by removing URLs, user mentions, hashtags, and emojis. The text is then preprocessed through tokenization, stopword removal, and lemmatization to standardize the input for modeling. The dataset is split into training and testing sets using a stratified approach to maintain the original sentiment distribution.

Modeling and Evaluation

After cleaning and preprocessing the tweets, machine learning models were trained to classify sentiment toward Apple and Google products. The models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, to ensure reliable and consistent sentiment predictions.

Model Comparison Results

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Baseline NB	6.549	5.963	4.043	3.921
Enhanced NB	6.748	5.909	5.192	5.406
Logistic Regression	6.455	5.461	6.143	5.649
Linear SVM	6.852	5.904	6.062	5.971
XGBoost	6.703	6.729	4.319	4.375

Linear SVM achieves the best overall performance, with the highest accuracy and Macro F1 score among all models.

Negative Sentiment Class Performance

Model	Negative Recall	Negative F1
Baseline NB	88	172
Enhanced NB	2.281	3.059
Logistic Regression	5.614	3.879
Linear SVM	4.825	4.435
XGBoost	702	1.270

Linear SVM provides the best balance for detecting negative sentiment, achieving the highest Negative F1 score, while Logistic Regression attains the highest Negative Recall.

Linear SVM Confusion Matrix

Actual	Negative	Neutral	Positive
Negative	Correctly identified negatives	Misclassified as neutral	Misclassified as positive
Neutral	Misclassified as negative	Correctly identified neutrals	Misclassified as positive
Positive	Misclassified as negative	Misclassified as neutral	Correctly identified positives

The confusion matrix shows that the Linear SVM performs best on the dominant neutral class, while still significantly improving the detection of negative sentiment compared to baseline models, though some confusion remains between positive and neutral tweets due to overlapping language.

Recommendations

The business should operationalize the Linear SVM by enabling automated crisis alerts, integrating negative tweets into customer service workflows, using sentiment dashboards to track trends during product launches, and establishing a continuous feedback loop to improve model performance.

Conclusion

This project applied NLP and machine learning techniques to analyze sentiment in Twitter data related to Apple and Google products. The results highlight how automated sentiment analysis can help organizations understand customer opinions and support informed business decisions.



THANK YOU