An NLP-driven study on social media sentiment trends

Twitter Sentiment Analysis: Apple vs. Google

Prepared by: Filda Kiarie

School: Moringa School

Email: filda.kiarie@student.moringaschool.com

Project Overview

Objective:

Classify tweets about Apple and Google as **positive, negative, or neutral** using NLP

Dataset:

CrowdFlower Twitter dataset (~9,000 tweets).

Dataset Source:

data world

Approach:

Data preprocessing

Feature engineering

Model development

Evaluation

Interpretation.

Data Preprocessing

Text Cleaning:

- ✓ Lowercasing
- ✓ Removing punctuation & special characters
- ✓ Stopword filtering

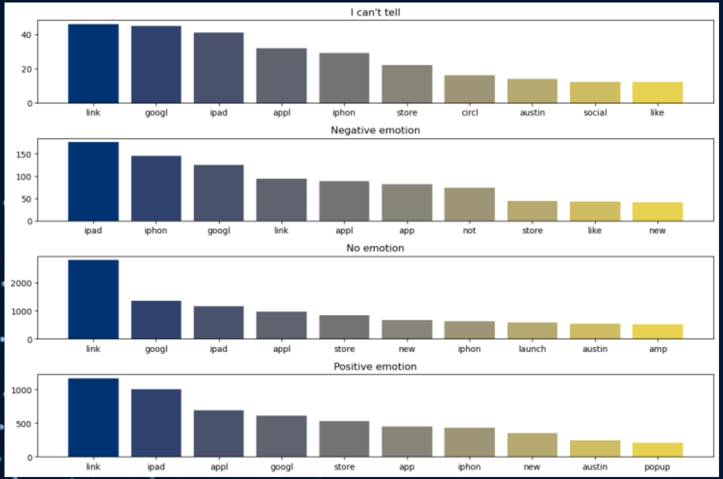
Feature Engineering:

- ✓ TF-IDF vectorization
- ✓ Tokenization

Data Exploration

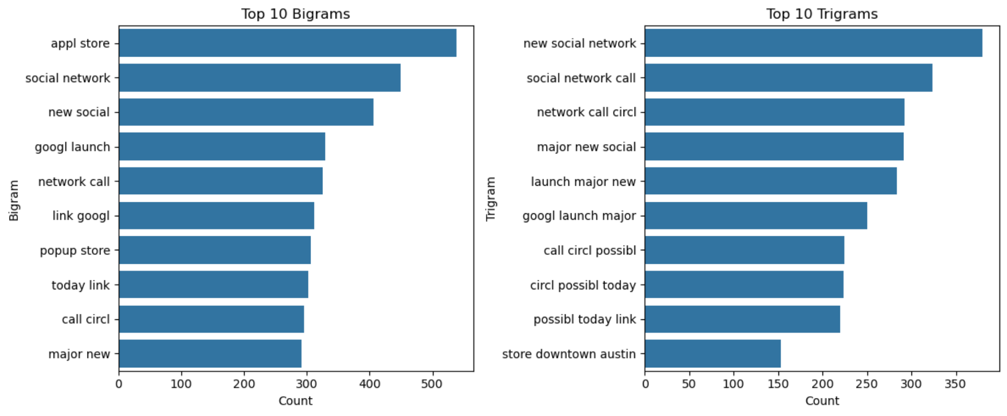
Frequent Words

This visual highlights the most commonly used words in the dataset, giving insight into key topics of discussion.



Bigrams & Trigrams:

This chart showcases the most frequent word combinations, helping identify popular phrases and trends.



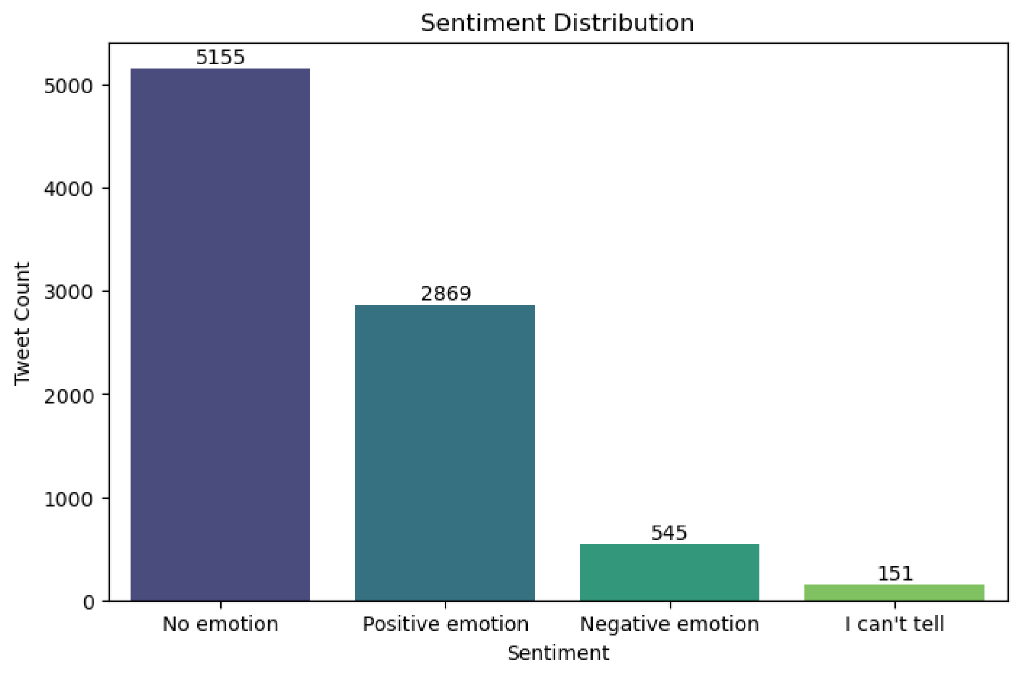
Sentiment Distribution

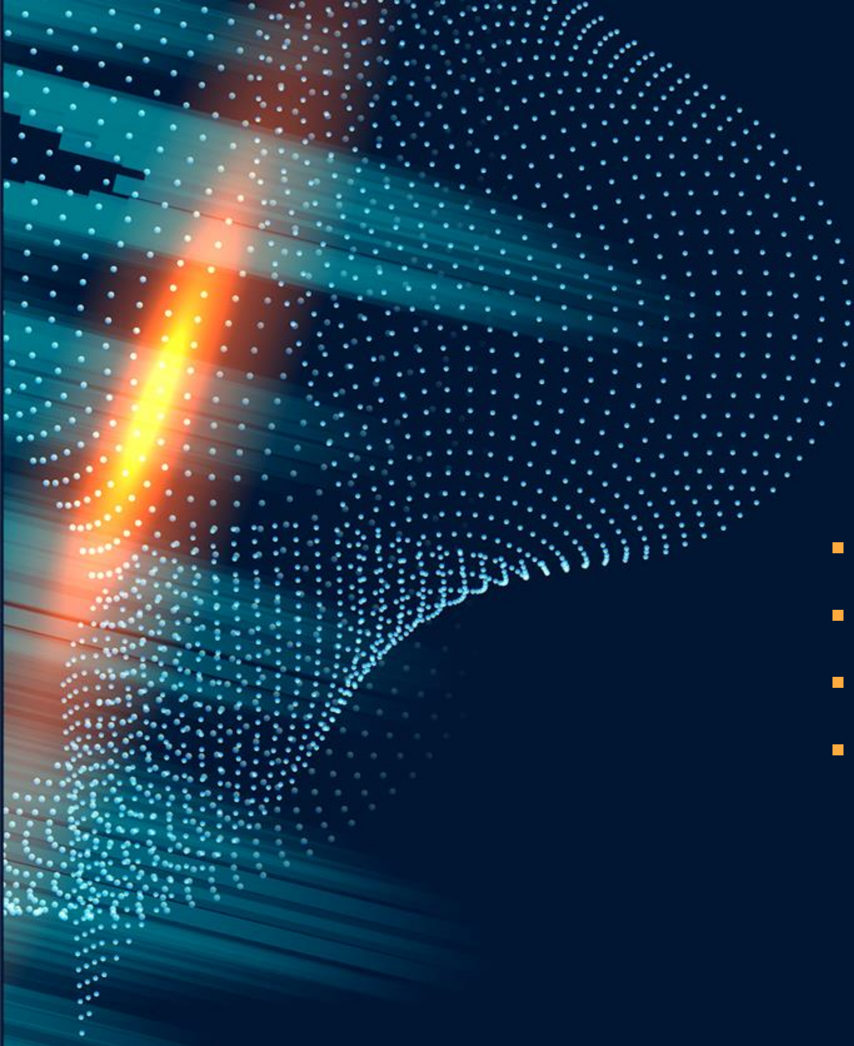
Objective: Categorizing tweets into positive, negative, and neutral sentiments.

•**Sentiment Distribution Chart:**

•This visualization displays the proportion of tweets classified as positive, negative, or neutral.

- **Key Takeaways:**
- Provides an overview of public perception.
- Useful for understanding sentiment trends for Apple vs. Google.





Model Selection

- Logistic Regression
- Support Vector Machine (SVM)
- XGBoost
- Evaluated using accuracy, precision, recall, F1-score, and SHAP analysis.

Model Performance Summary

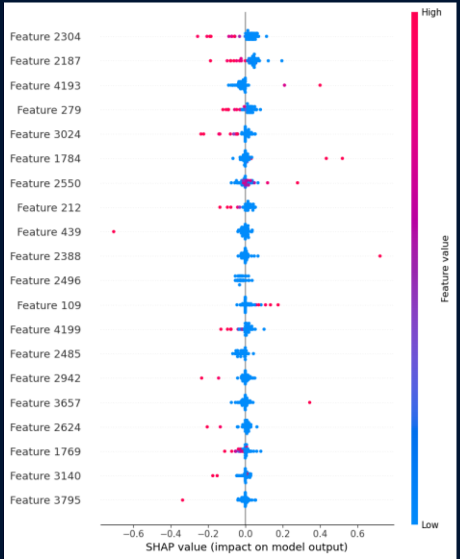
Model	Accuracy	Precision	Recall	F-score
Logistic Regression	96%	High	High	High
SVM(Best)	98%	Best	Best	Best
XGBoost	92%	Moderate	Moderate	Moderate

Key Takeaways:

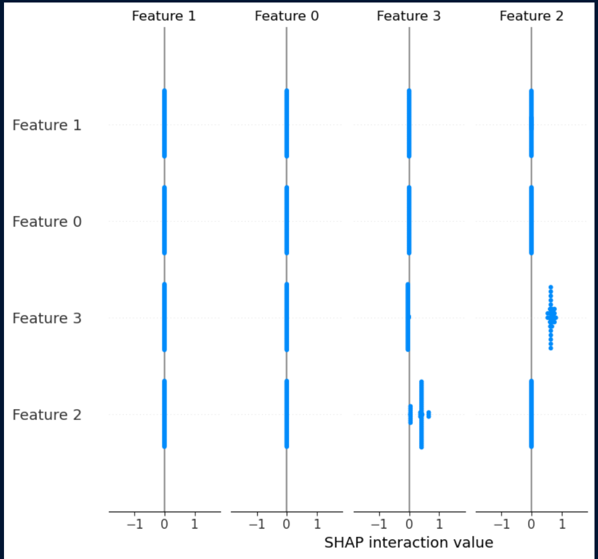
- **SVM outperformed all models** with the highest accuracy and balanced classification.
- **Logistic Regression** remains valuable for interpretability.
- **XGBoost** was slightly weaker but still useful.

SHAP Interpretability:

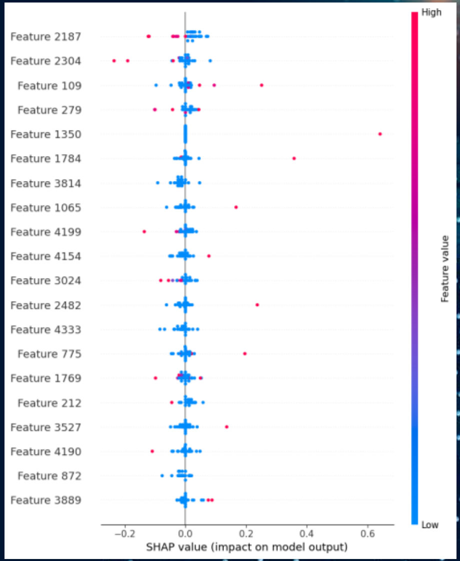
Logistic Regression SHAP Output: Identified key words contributing to sentiment classification.



SVM SHAP Output: Confirmed strong feature importance consistency.



XGBoost SHAP Output: Highlighted word influence but with more variance.



Conclusion: SVM consistently had **more reliable feature explanations** for decision-making.

Key Insights

- **SVM is the most accurate model (98%)** and balances all sentiment classes effectively.
- **Logistic Regression (96%)** offers strong interpretability for tracking sentiment trends.
- **XGBoost (92%)** struggled slightly with class imbalances but remains a useful alternative.
- **Misclassification was highest between neutral and negative sentiment classes.**

THANKS!

Do you have any questions?
filda.kiarie@student.moringaschool.com

Git Username: **FildaKim**