android

CLASSIFYING A TWEET'S SENTIMENT BASED ON ITS CONTENT

By DS-PT II, Group 6

MORINGA
Discover · Grow · Transform

OUR TEAM

Charity Mwangangi
Keith Tongi
Edgar Muturi
Jacob Abuon
Edna Maina
Kevin Karanja

PURPOSE OF THE PROJECT

The purpose of this project is to classify tweets mentioning different Apple and Google brands into sentiment categories, in order to analyse consumer attitudes and compare how sentiment varies across brands.

This helps demonstrate the use of Natural Language Processing for social media monitoring and brand analysis.

THE PROCESS (OSEMN)

Obtain – Description of our dataset & source.
Scrub – Description of cleaning steps.
Explore – Key trends & insights.
Model – Algorithms used.
Interpret – Findings & applications.

THE PROCESS (CTD...)

Obtain

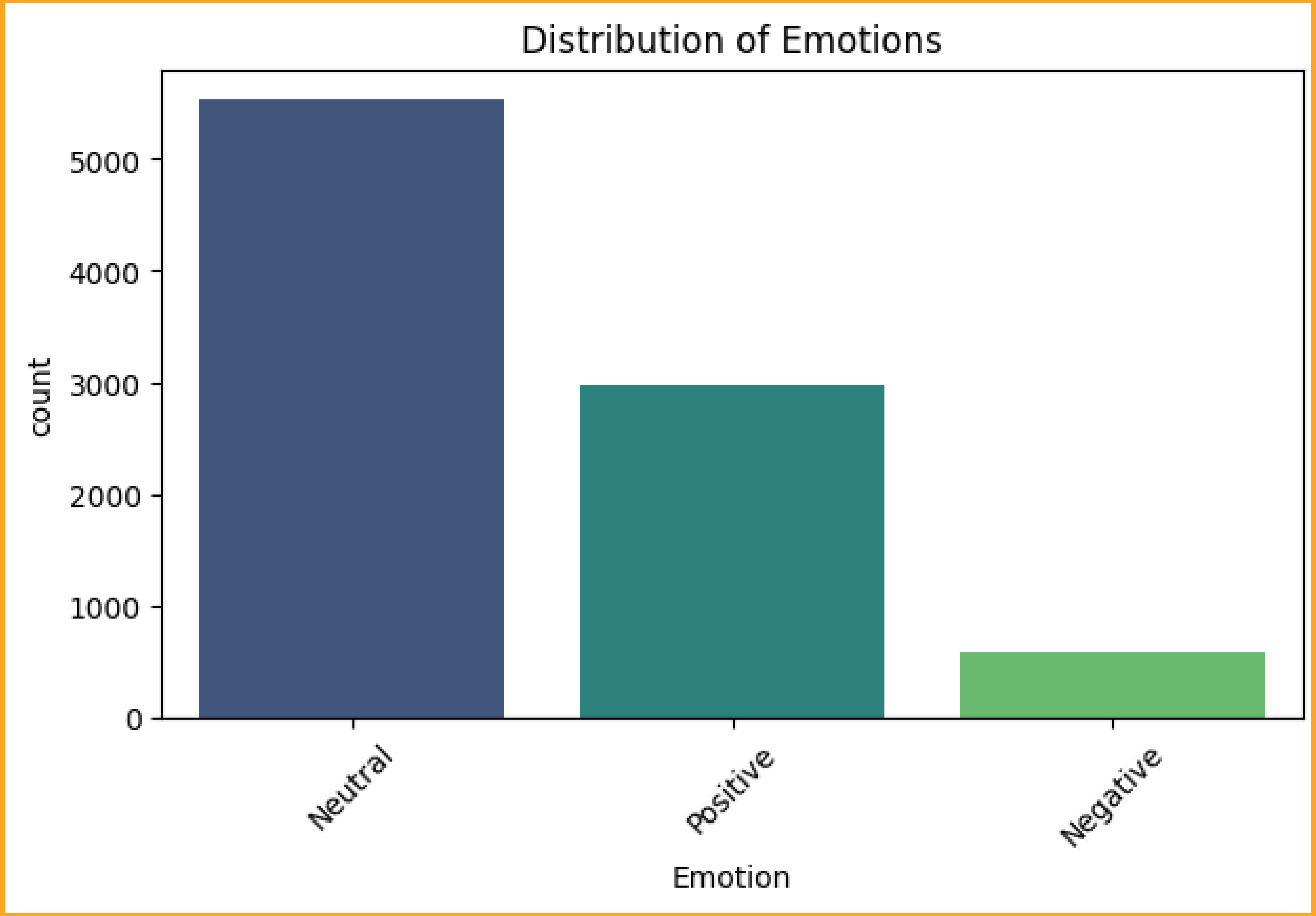
- Dataset origin - CrowdFlower.
- Dataset name - Tweet_data.csv (https://data.world/crowdflower/brands-and-product-emotions)

Scrub

- Text Pre-processing - Remove stopwords, Normalise text, Lemmatisation.
- Handling missing values or imbalances.

THE PROCESS (CTD...)

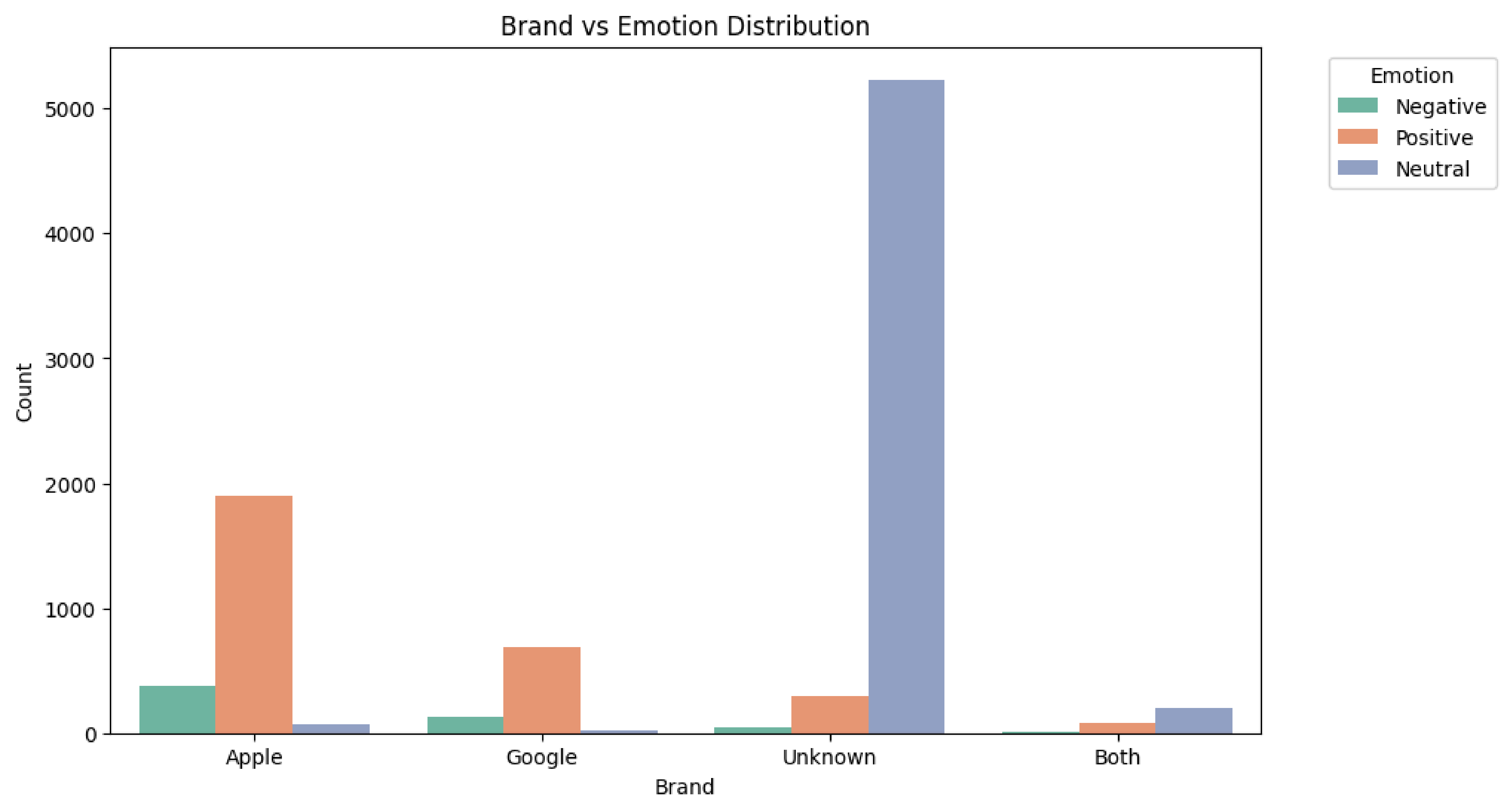
Exploratory Data Analysis



This chart reveals that a significant amount of tweets in our dataset are neutral, followed by positive tweets.
Negative tweets are less than a thousand.

THE PROCESS (CTD...)

Exploratory Data Analysis

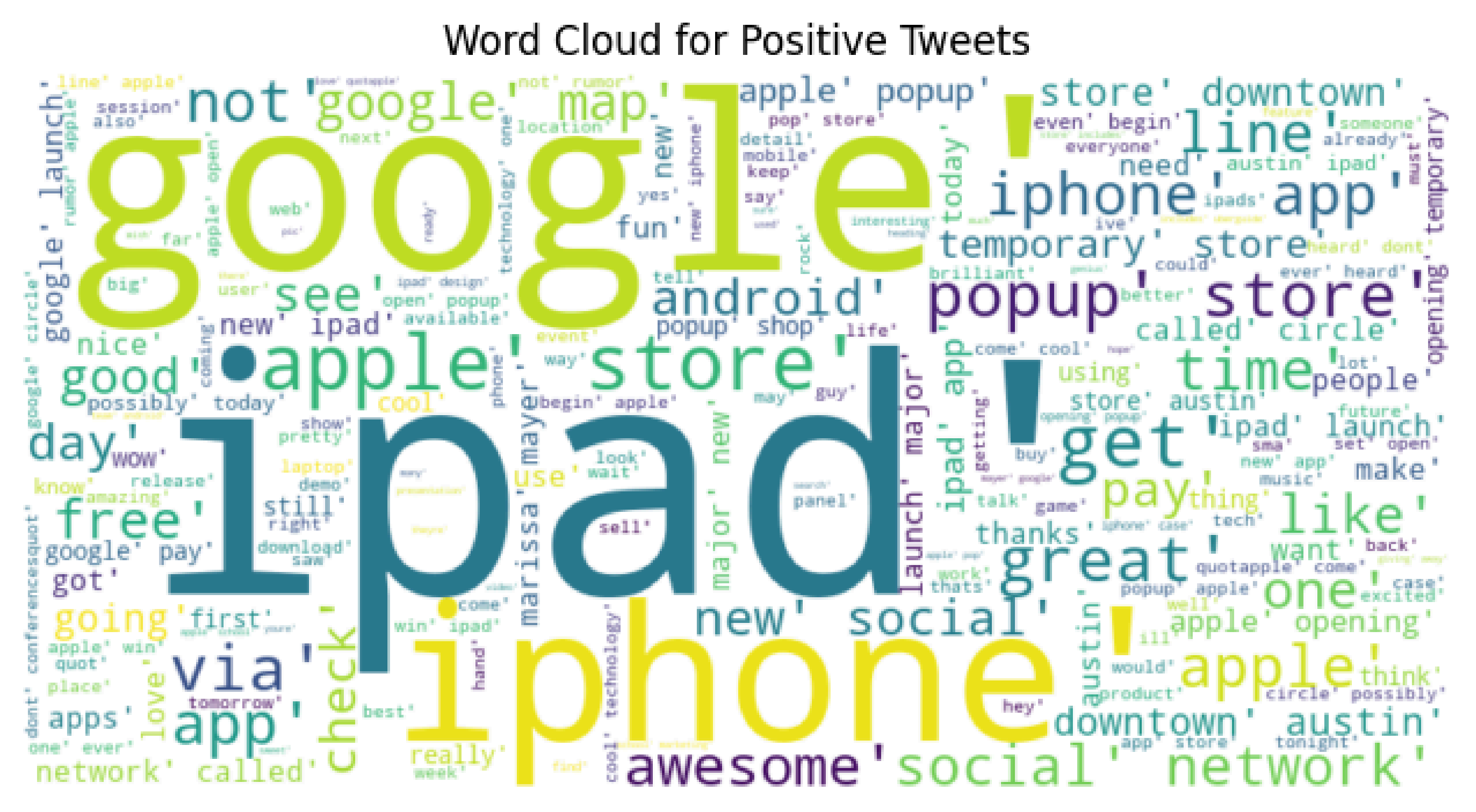


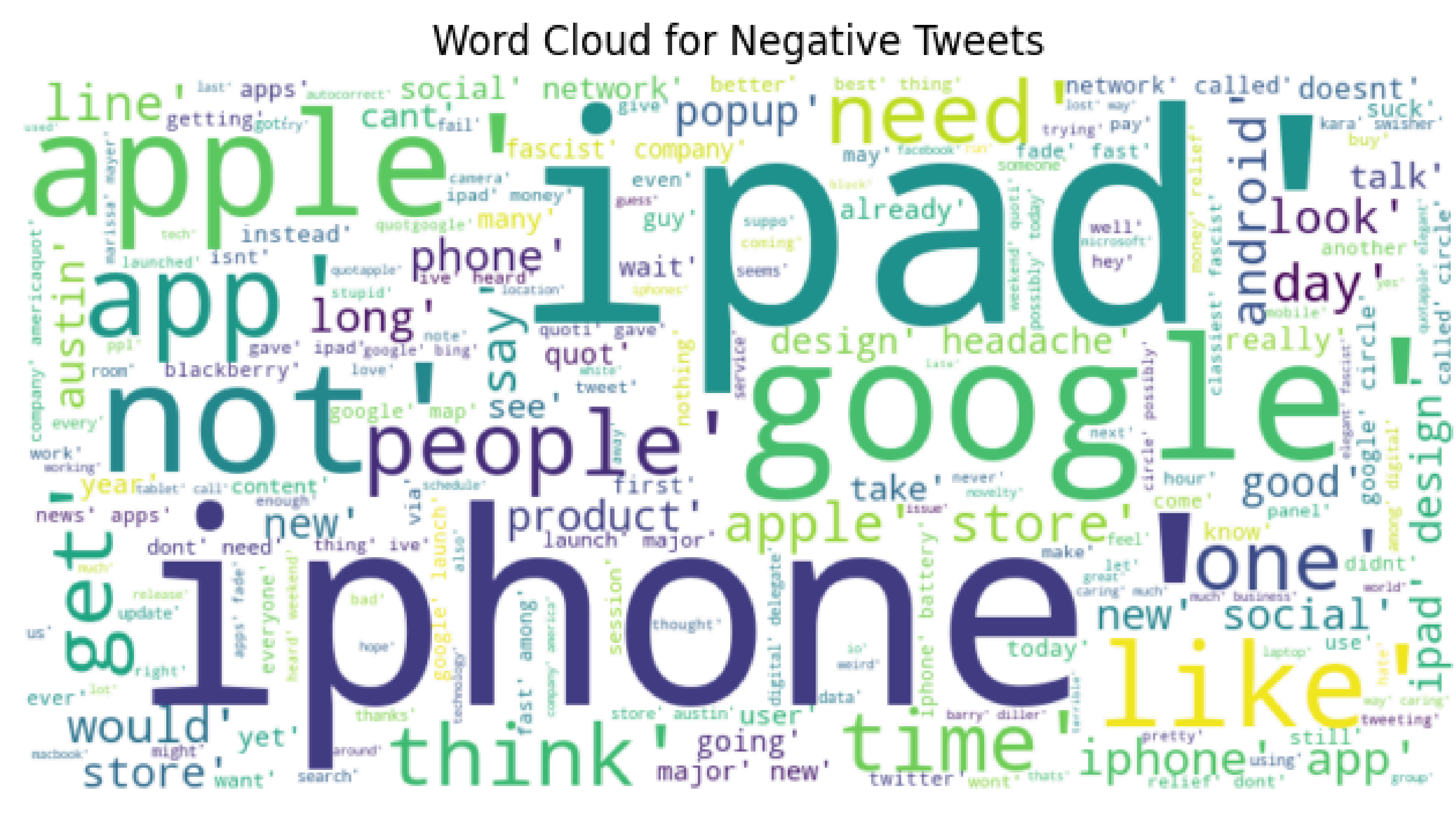
This chart reveals that both products possess positive, neutral and negative sentiment.

A bigger proportion addresses apple products.

THE PROCESS (CTD...)

Exploratory Data Analysis

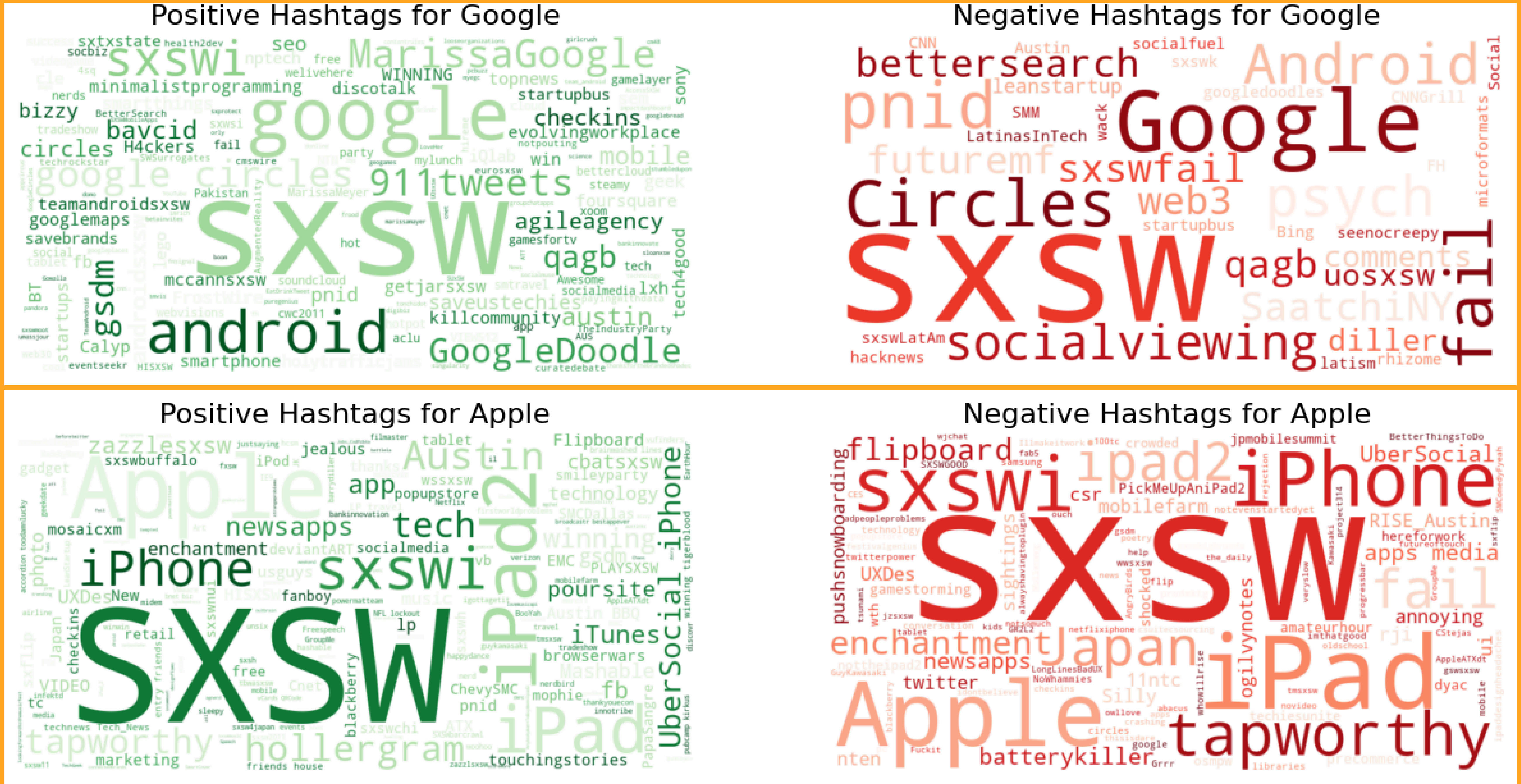




Wordclouds for positive and negative tweets.

THE PROCESS (CTD...)

Exploratory Data Analysis


Positive Hashtags for Google

Negative Hashtags for Google

Positive Hashtags for Apple

Negative Hashtags for Apple

Wordclouds for positive and negative hashtags.

MODELLING APPROACH

- Algorithms tested include Logistic Regression, Naive Bayes, and Random Forest Classifier.
- Vectorizers used include CountVectorizer and TF-IDF.
- Baseline results.

MODELLING AND METHODOLOGY

Aspect	Details
Baseline Model	Binary Logistic Regression (positive vs. negative)
Multiclass Models Tested	Logistic Regression, LinearSVC, Multinomial Naive Bayes, Random Forest, Gradient Boosting
Feature Extraction	CountVectorizer, TF-IDF Vectorizer
Hyperparameters Tuned	Vectorizers: n-gram, min_df; Logistic Regression: C; LinearSVC: C; Naive Bayes: alpha; Random Forest: n_estimators, max_depth; Gradient Boosting: n_estimators, learning rate
Validation Strategy	StratifiedKFold (5 folds)
Primary Metric	Macro F1 (robust for imbalanced data)
Accuracy, ROC-AUC, Weighted AUC	Accuracy, ROC-AUC, Weighted AUC

RESULTS AND INSIGHTS

Class	Precision	Recall	F1-Score	ROC-AUC
Negative (0)	0.43	0.37	0.4	0.85
Positive (1)	0.76	0.74	0.75	0.76
Neutral (2)	0.57	0.61	0.59	0.74
Macro Average	–	–	**0.5671 (CV F1)**	**0.78 (Macro AUC)**
Weighted Average	–	–	–	**0.75 (Weighted AUC)**

- Best Model: Logistic Regression with TF-IDF vectorization.

LIMITATIONS

- **Class Imbalance** - Negative tweets are underrepresented, leading to weaker recall and F1 performance.
- **Lexical Focus** - Models rely on word counts (TF-IDF, n-grams), which may miss context, sarcasm, or slang.
- **Generalisability** - Performance is tied to the training dataset; results may not transfer well to other topics or timeframes.
- **Feature Limitations** - No use of semantic embeddings (e.g., Word2Vec, BERT) or sentiment lexicons that could capture deeper meaning.
- **Evaluation Constraints** - Cross-validation improves reliability, but a larger, more balanced dataset is needed for robust generalization.

CONCLUSION

Business recommendations from our insights:

- **Boost Support** - Faster response, chatbots, FAQs
- **Use Positives** - Testimonials, campaigns, loyalty
- **Guide R&D** - Fix pain points, align with feedback
- **Track Competitors** - Compare strengths, refine positioning
- **Crisis Readiness** - Real-time monitoring, rapid PR response

THANK YOU QUESTIONS?