Twitter Sentiment Analysis on Tech Products

PRESENTED BY: **GROUP 3**

MEET OUR TEAM

Wamuyu Gitonga – Team Lead

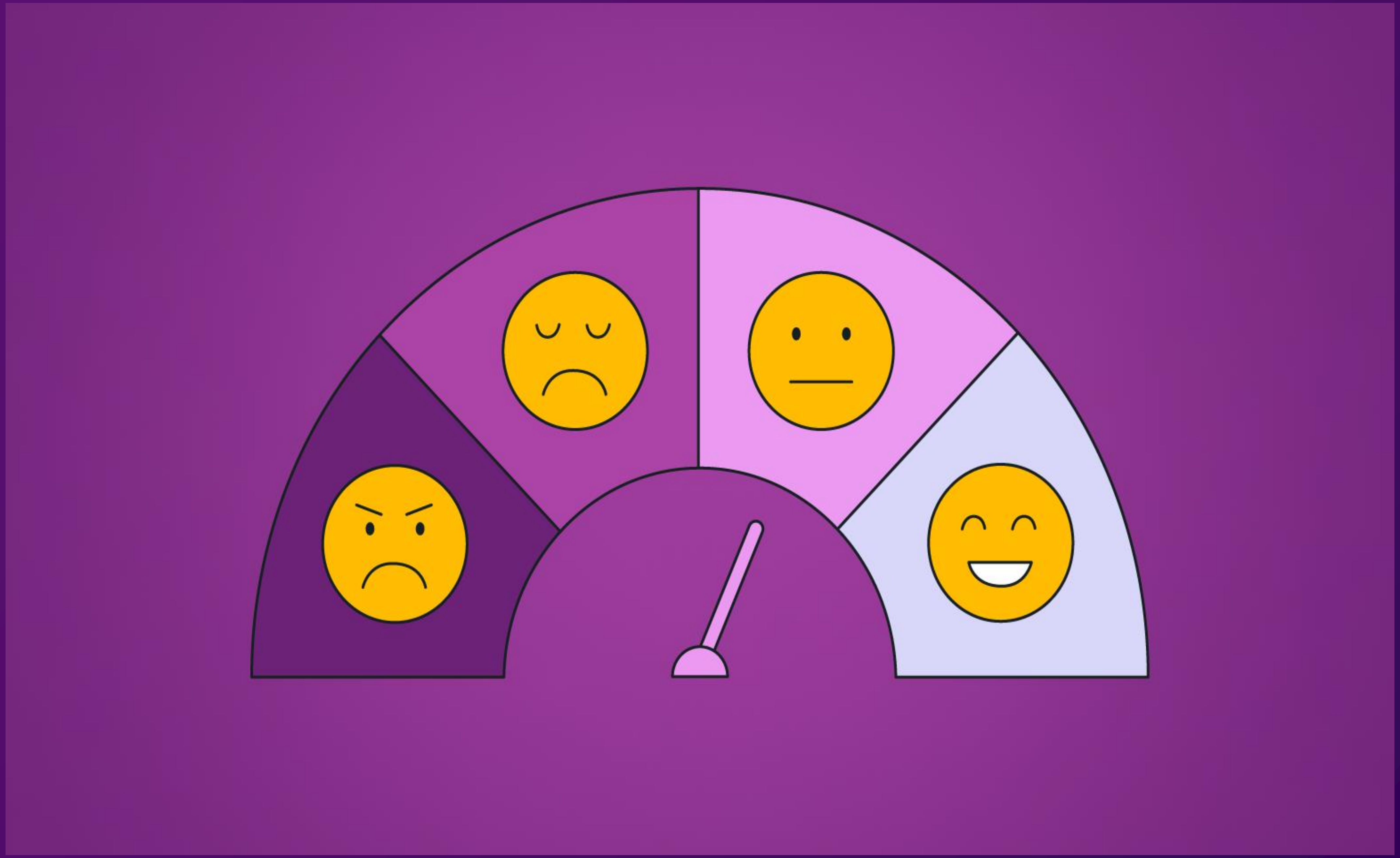
Vanessa Sandra

John Njogu

Natalie Wanjiku

Leejay Mwakireti

Lillian Wangui



AGENDA

- Data Source
- Objective
- Dataset Overview
- Data Quality
- Descriptive Analysis
- Trends & Categories
- Model Performance
- Key Insights
- Recommendations
- Conclusion

DATA SOURCE

The dataset comes from CrowdFlower via [data.world](https://data.world/crowdflower/brands-and-product-emotions/workspace/file?filename=judge-1377884607_tweet_product_company.csv).

This dataset is highly suitable because the text of Tweets directly reflects customer opinions. Analyzing these Tweets allows stakeholders (product managers, marketing teams) to understand public sentiment efficiently without manually reading thousands of messages.

OBJECTIVES

1) Analyze Twitter conversations about Apple, Google, and Android products using natural language processing (NLP) techniques

2) Classify tweets into positive, negative and neutral sentiment patterns

3) Build Machine Learning (ML) models to classify sentiments automatically

4) Extract insights to support businesses, marketers, and stakeholders in the tech industry to track consumer sentiment, benchmark against competitors, and drive data-informed decisions

DATASET OVERVIEW

1) The CSV file contains 9093 rows of tweets and 3 columns

2) Columns include: tweet text, product mentioned, sentiment label

3) Focus: Sentiment directed at tech brands/products

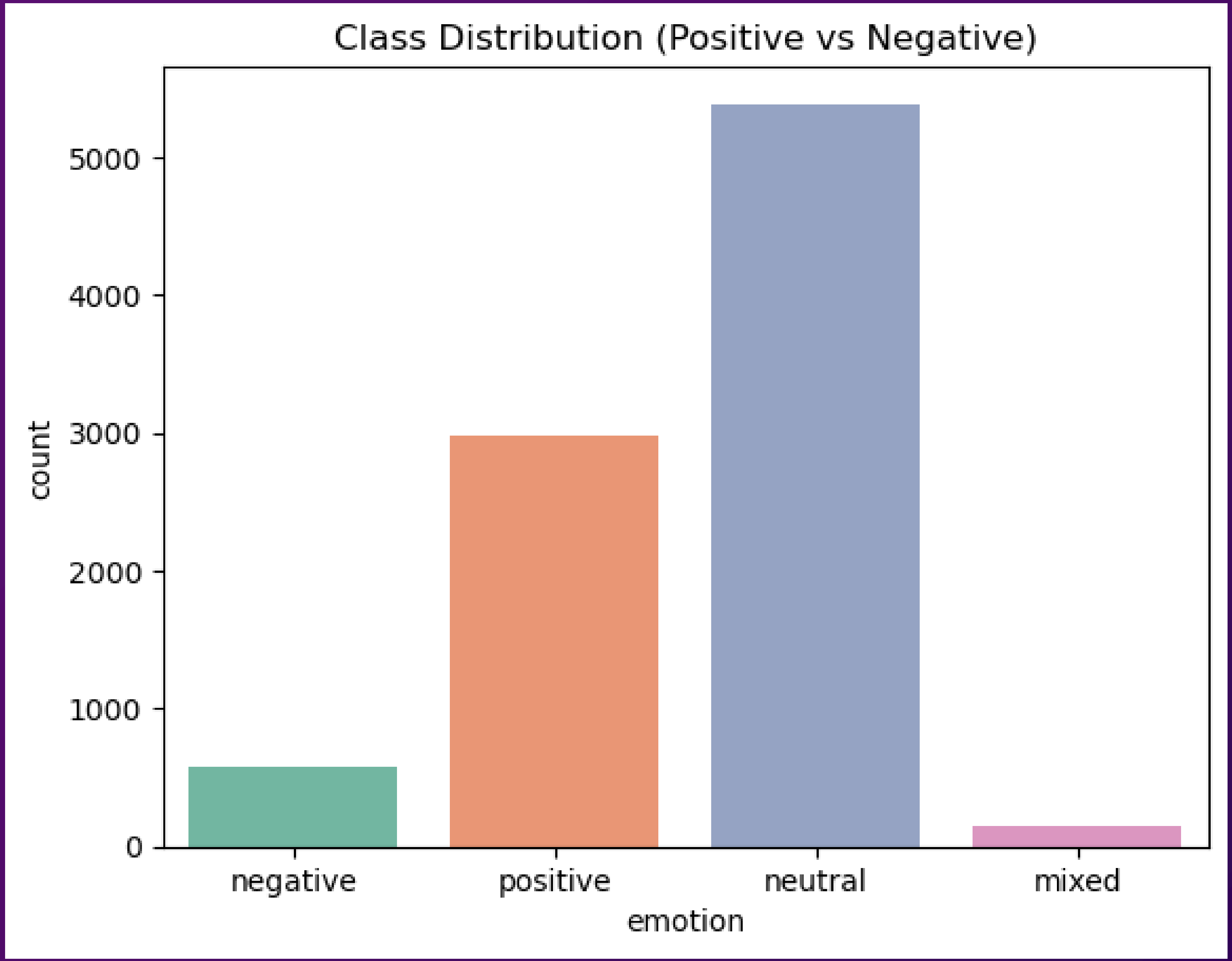
DATASET QUALITY CHECKS

1) Checked for missing values and duplicated records

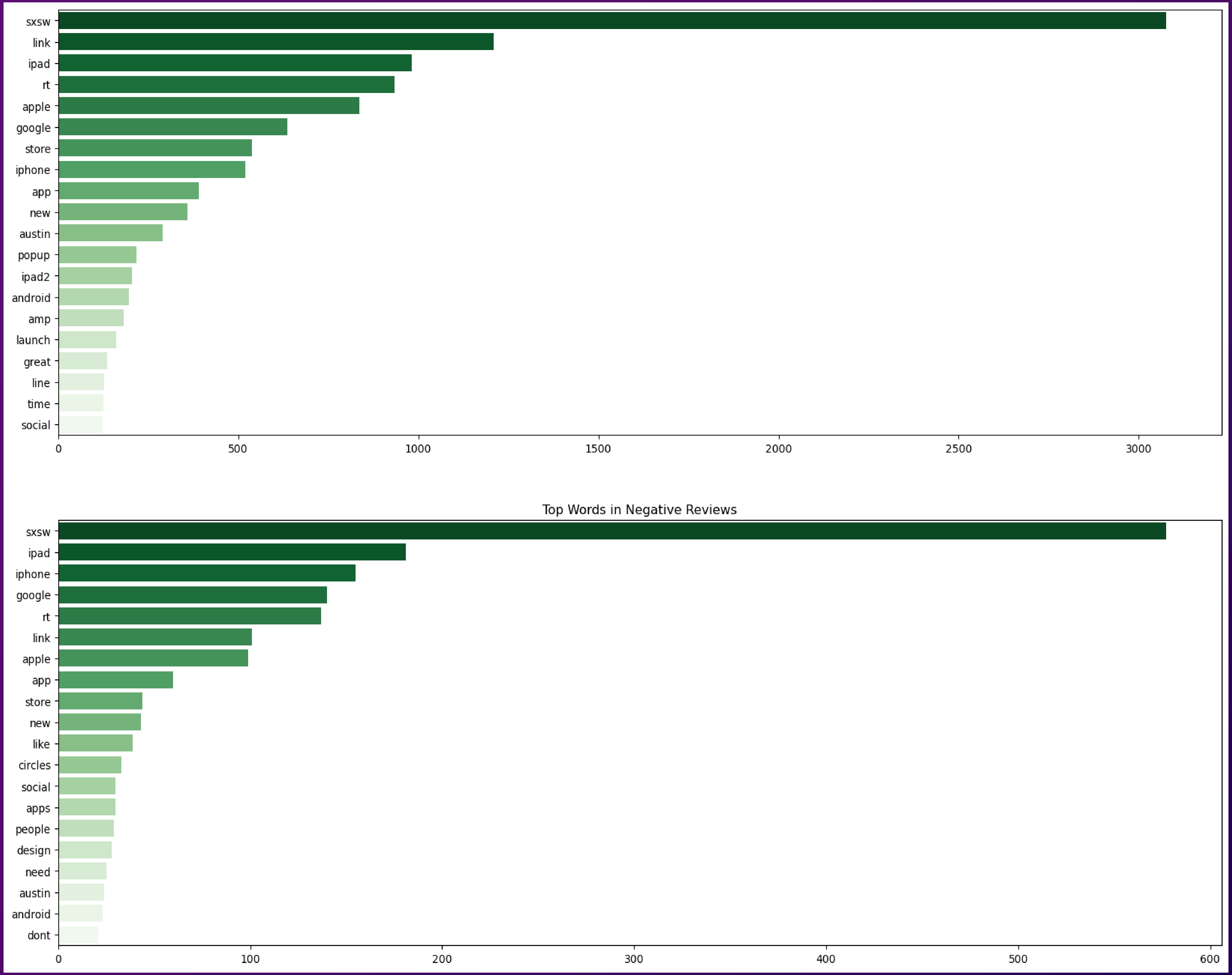
2) Validated sentiment categories (Positive, Negative, Neutral) and standardization

3) Cleaned noisy text for analysis

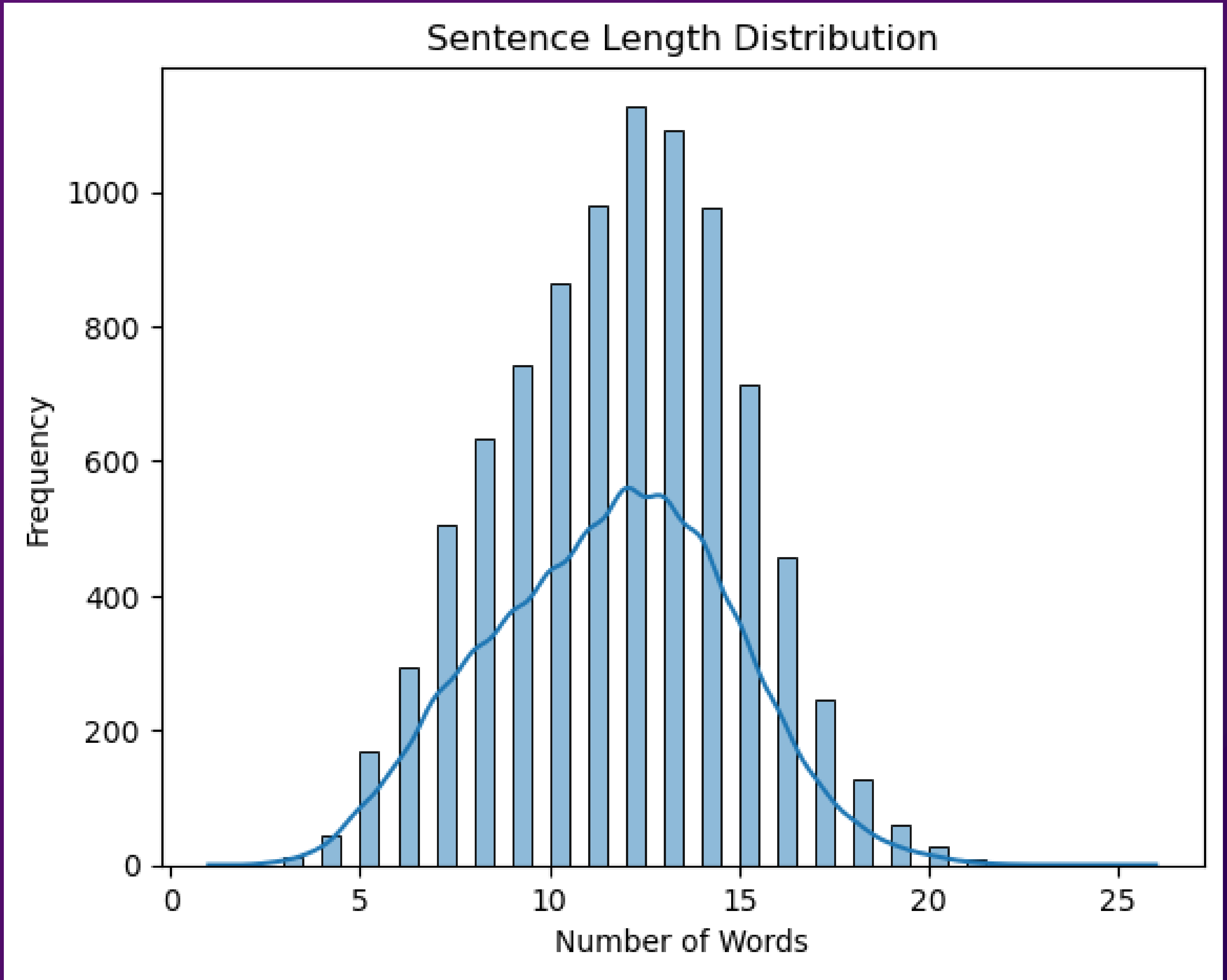
EXPLORATORY DATA ANALYSIS



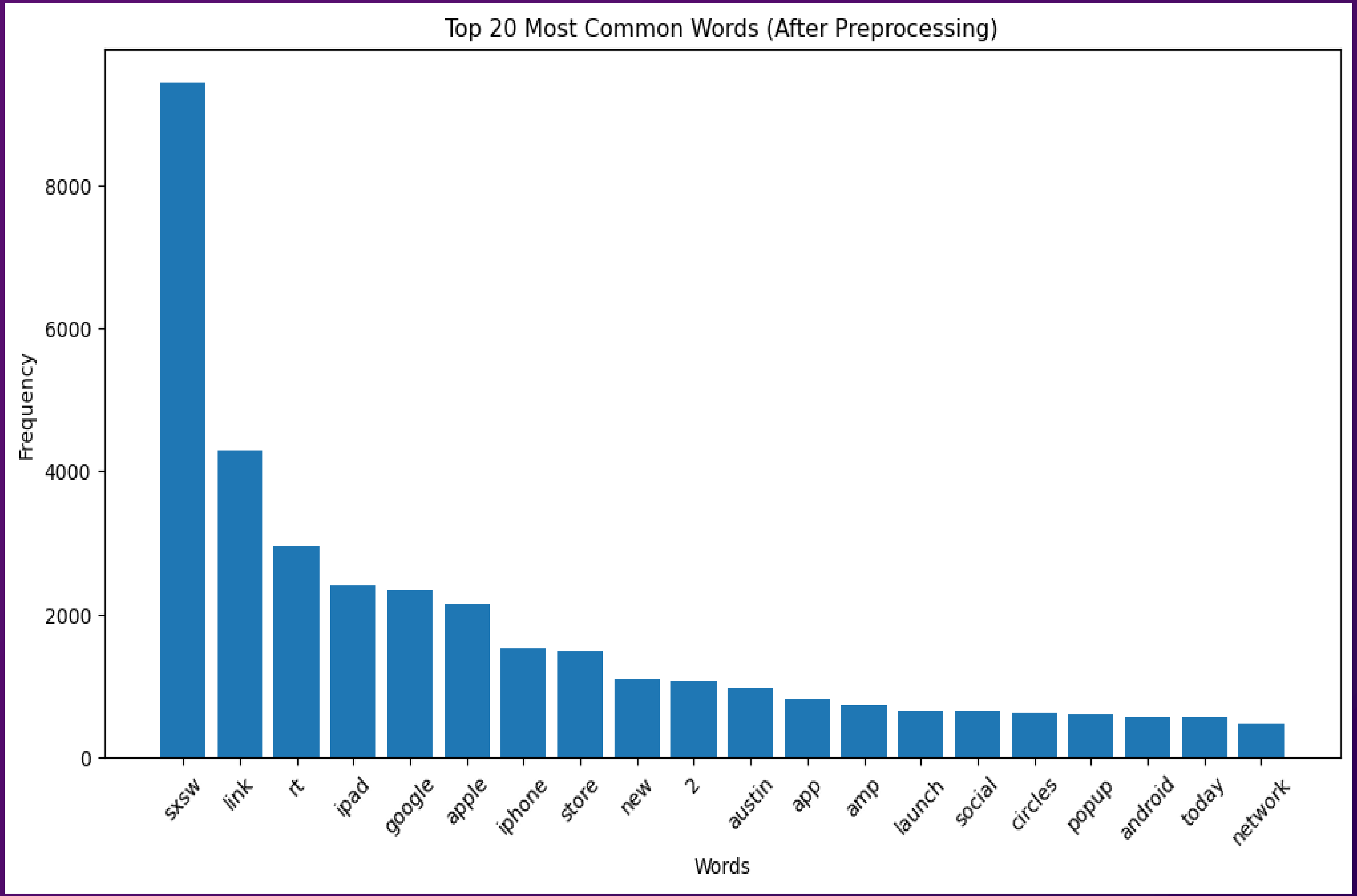
Shows the distribution of sentiments



Shows the distribution of words in Positive and Negative Reviews



Shows the sentence length distribution



Shows the most common words used after preprocessing

MODEL TRAINING

We trained our models to correctly classify whether a tweet is negative or positive. The models used were:
- Logistic Regression
- SVM
- Random Forest
- XGBoost
- Gradient Descent

TF-IDF to vectorize our data, and logistic regression to classify it.
SMOTE applied to handle class imbalance

MODEL EVALUATION

This table compares five different machine learning models across five key evaluation metrics.

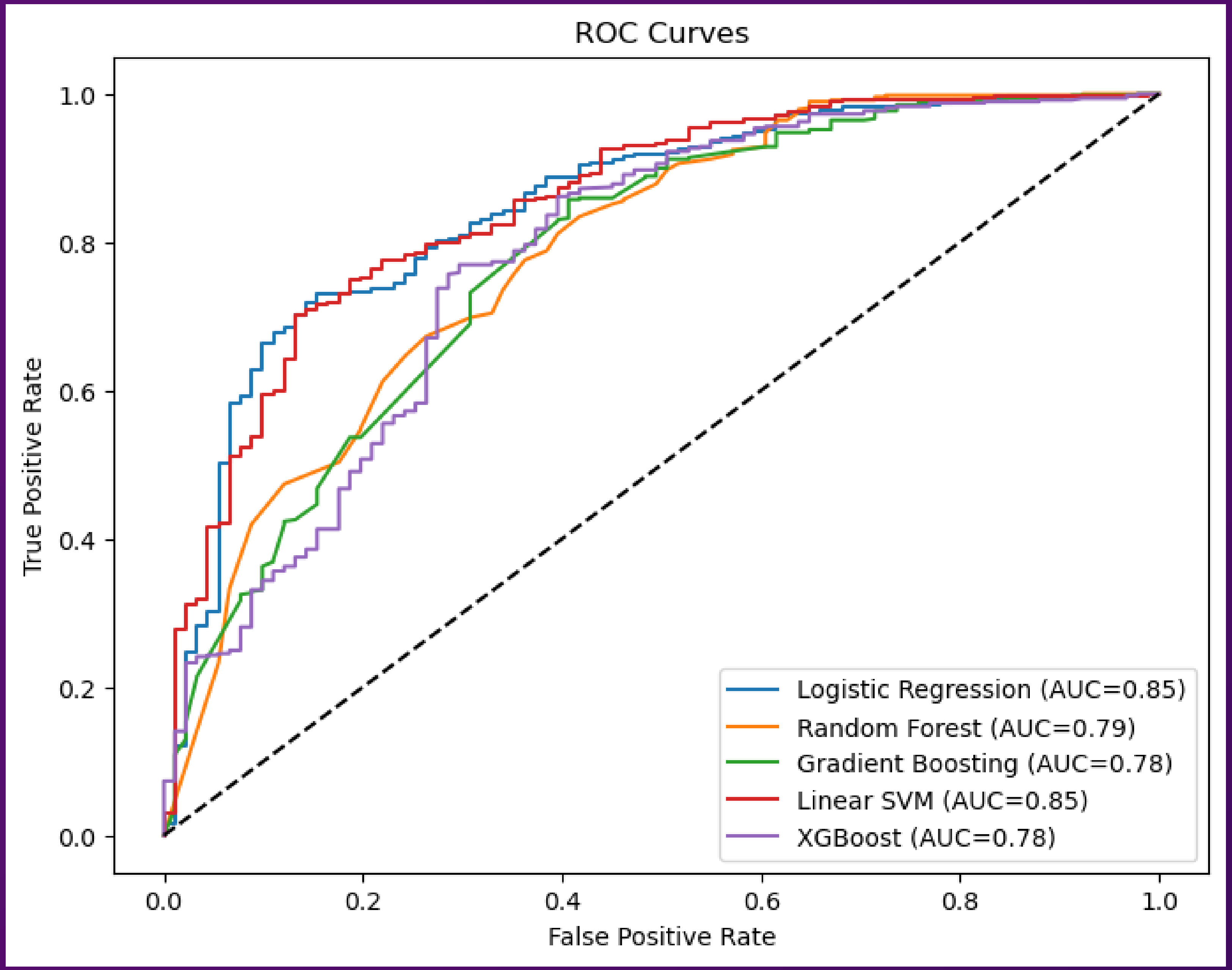
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	**0.847**	0.847	**0.998**	0.916	0.848
Random Forest	**0.882**	0.880	**0.996**	**0.934**	0.785
Gradient Boosting	0.861	0.864	0.990	0.923	0.778
Linear SVM	0.877	**0.888**	0.977	0.930	**0.851**
XGBoost	0.866	0.875	0.981	0.925	0.779

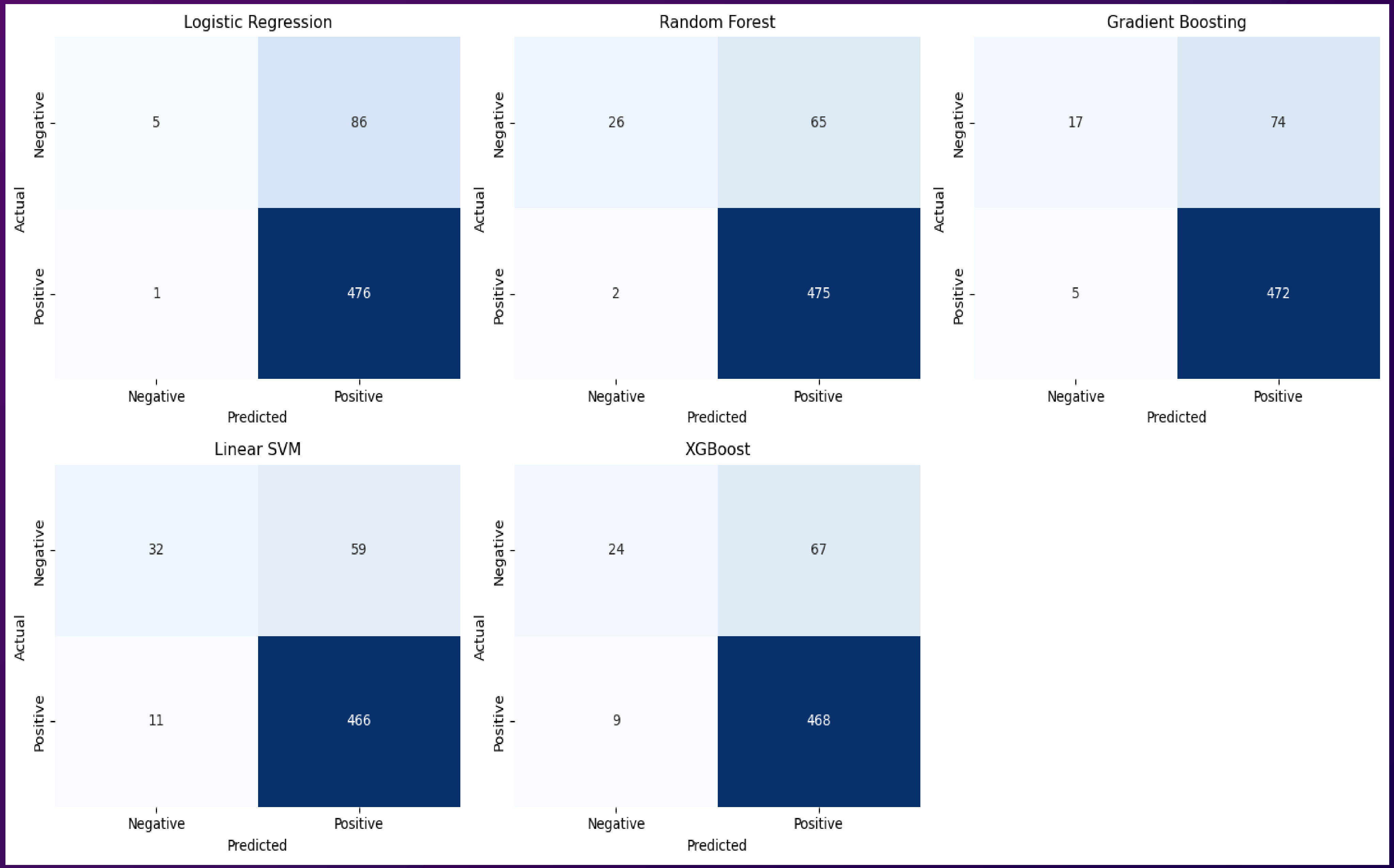
Summary and Recommendation

Best Overall Model: Random Forest, due to its highest Accuracy and F1-score, which best balances the trade-off between false positives and false negatives.

Best for High-Certainty Predictions: Linear SVM, if your priority is minimizing false positives (i.e., you want to be very sure a prediction is positive when you make it).

Key Observation: All models have exceptionally high Recall (>0.97). This means they are all very good at capturing nearly all positive instances in the data. The primary differences between them lie in how many false positives they generate (reflected in Precision).


ROC Curves
True Positive Rate
False Positive Rate
Logistic Regression (AUC=0.85)
Random Forest (AUC=0.79)
Gradient Boosting (AUC=0.78)
Linear SVM (AUC=0.85)
XGBoost (AUC=0.78)


Logistic Regression
5
86
1
476
Random Forest
26
65
2
475
Gradient Boosting
17
74
5
472
Linear SVM
32
59
11
466
XGBoost
24
67
9
468
Actual
Predicted
Negative
Positive

KEY INSIGHTS

1. Random Forest is the best model overall
2. Linear SVM performs strongly in ROC-AUC
3. Sentiment classification is highly accurate (>85%)
4. Tech brands have strong positive sentiment, especially Apple

RECOMMENDATIONS

1. General Sentiment Trends – Track overall positive vs. negative sentiment (e.g., “60% positive, 40% negative”) and address spikes in negativity with responsive campaigns.

2. Positive Drivers – Amplify what users already love (e.g., Google Pixel’s camera, iPhone’s ecosystem) through ads, influencer marketing, and social proof.

3. Negative Pain Points – Identify top complaints (e.g., iPhone battery life, Android bloatware) and directly counter them with messaging, feature highlights, or improved support.

4. Competitive Benchmarking – Compare sentiment across Apple vs. Google, iPhone vs. Android, and use competitor weaknesses as positioning opportunities.

5. Actionable Monitoring – Build a live dashboard to track sentiment shifts after launches or campaigns, ensuring marketing actions are data-driven and adaptive.

CONCLUSION

1. Dataset provides valuable insights on customer sentiment
2. ML models achieve high accuracy and reliability
3. Findings can guide marketing, product strategy, and customer engagement
4. Next step: Implement sentiment monitoring pipeline

THANK YOU