Tweet Sentiment Analysis

Understanding emotions and opinions in social media posts

INTRODUCTION

This project leverages Natural Language Processing (NLP) techniques to analyze public sentiment toward Apple and Google products. By using a dataset of 9,093 tweets, each labeled as positive, negative, or neutral, we aim to develop predictive models capable of accurately classifying sentiment.

The primary objective is to create a proof-of-concept sentiment analysis system that not only predicts sentiment accurately but also provides interpretable insights for decision-making. This project highlights a structured, end-to-end NLP pipeline, demonstrating our proficiency in data preprocessing, feature engineering, model development, evaluation, and business-oriented interpretation of results.

Stakeholders and Business Value

Stakeholders

- Tech Companies: Track customer satisfaction, detect issues, and prioritize product improvements.
- Social Media Analysts & Data Scientists: Monitor trends and generate actionable insights.
- Investors & Strategists: Inform investment decisions and competitive strategies.
- Researchers: Study NLP, social media trends, and consumer behavior.
- Customers: Benefit indirectly through improved products, services, and brand experience.

Business Value

- Identify product satisfaction and potential issues.
- Optimize marketing, engagement, and communication strategies.
- Guide strategic planning and investment decisions.
- Support research in NLP and social media analytics.
- Enhance customer experience through feedback-driven improvements.

Dataset Properties & Limitations

Dataset Composition

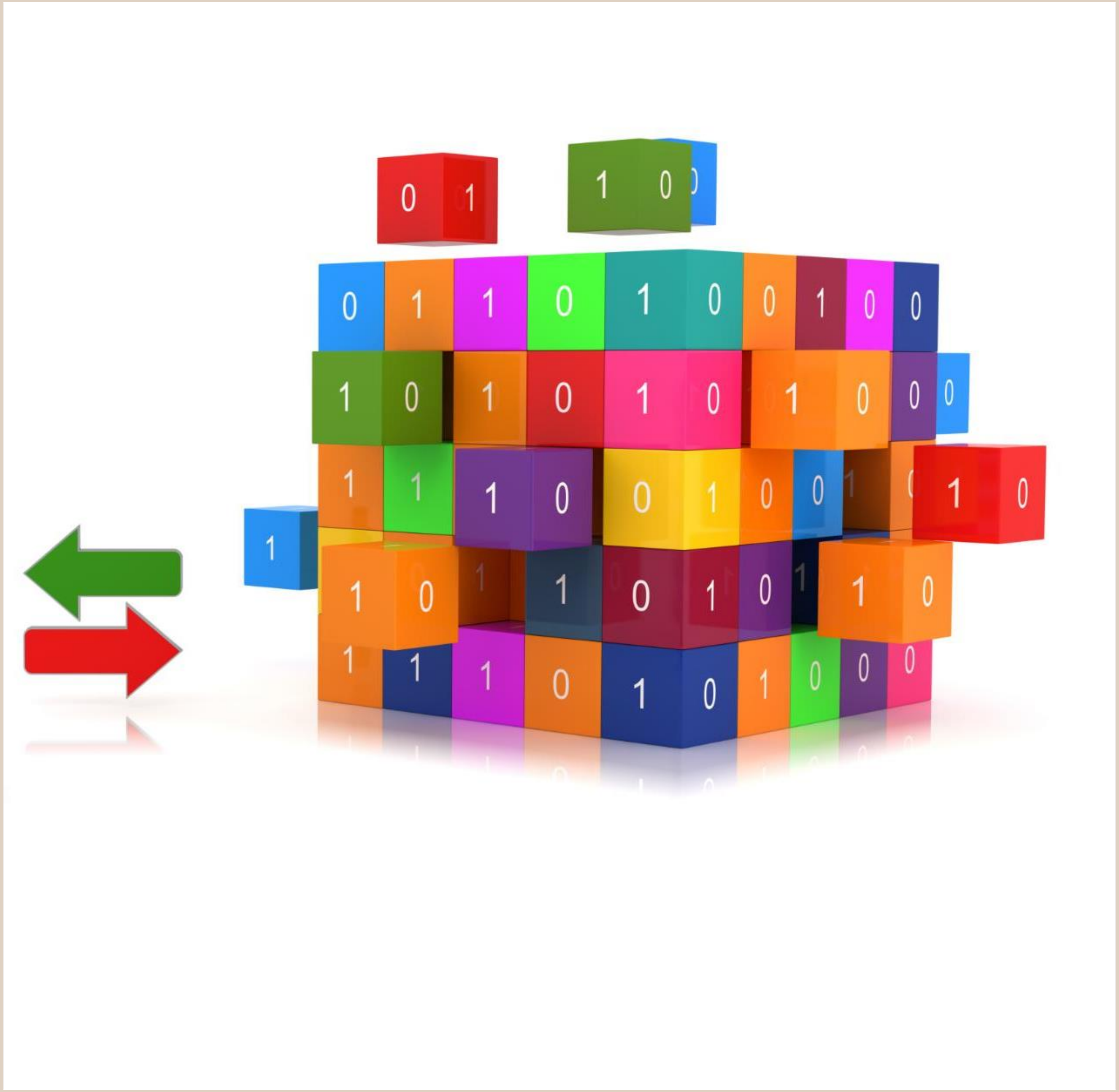
Dataset contains raw tweets with informal language and emojis labeled as positive, negative, or neutral sentiments.

Use and Suitability

Dataset is suitable for supervised learning and reflects real-world social media sentiment accurately.

Limitations and Challenges

Limitations include language bias, slang complexity, class imbalance, and outdated data requiring preprocessing and tuning.



Data Cleaning & Preprocessing

Removing Missing and Duplicate Data

Cleaning removed missing and duplicate tweets to ensure accuracy and relevance of the dataset.

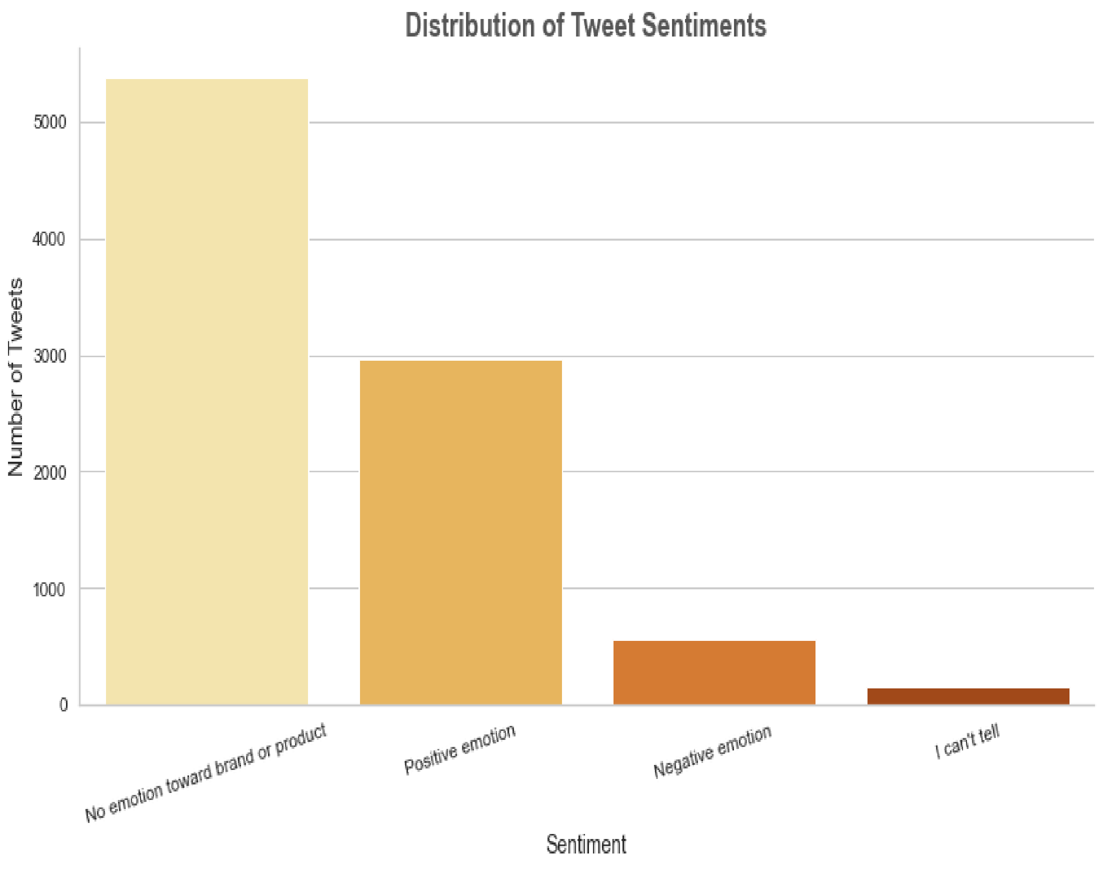
Handling Missing Brand References

Missing brand or product references were assigned 'Unknown' to retain all tweets for analysis.

Ensuring Data Quality for NLP

High-quality cleaned data supports robust feature engineering and improves classification model performance.

Sentiment Distribution & Insights



Sentiment Distribution

Neutral tweets dominate (~5,000), with positive tweets around 2,500–3,000, and negative tweets fewer at 500–700.

Word Cloud Analysis

Frequent words include SXSW, iPhone, iPad, and Google, highlighting popular topics in tweets.

Sentiment Language Differences

Positive tweets use words like 'great' and 'awesome', while negative tweets include 'headache' and 'fascist'.

Tweet Length Variation

Negative tweets tend to be longer in length compared to positive and neutral tweets.

Binary Classification

Binary Classification Focus

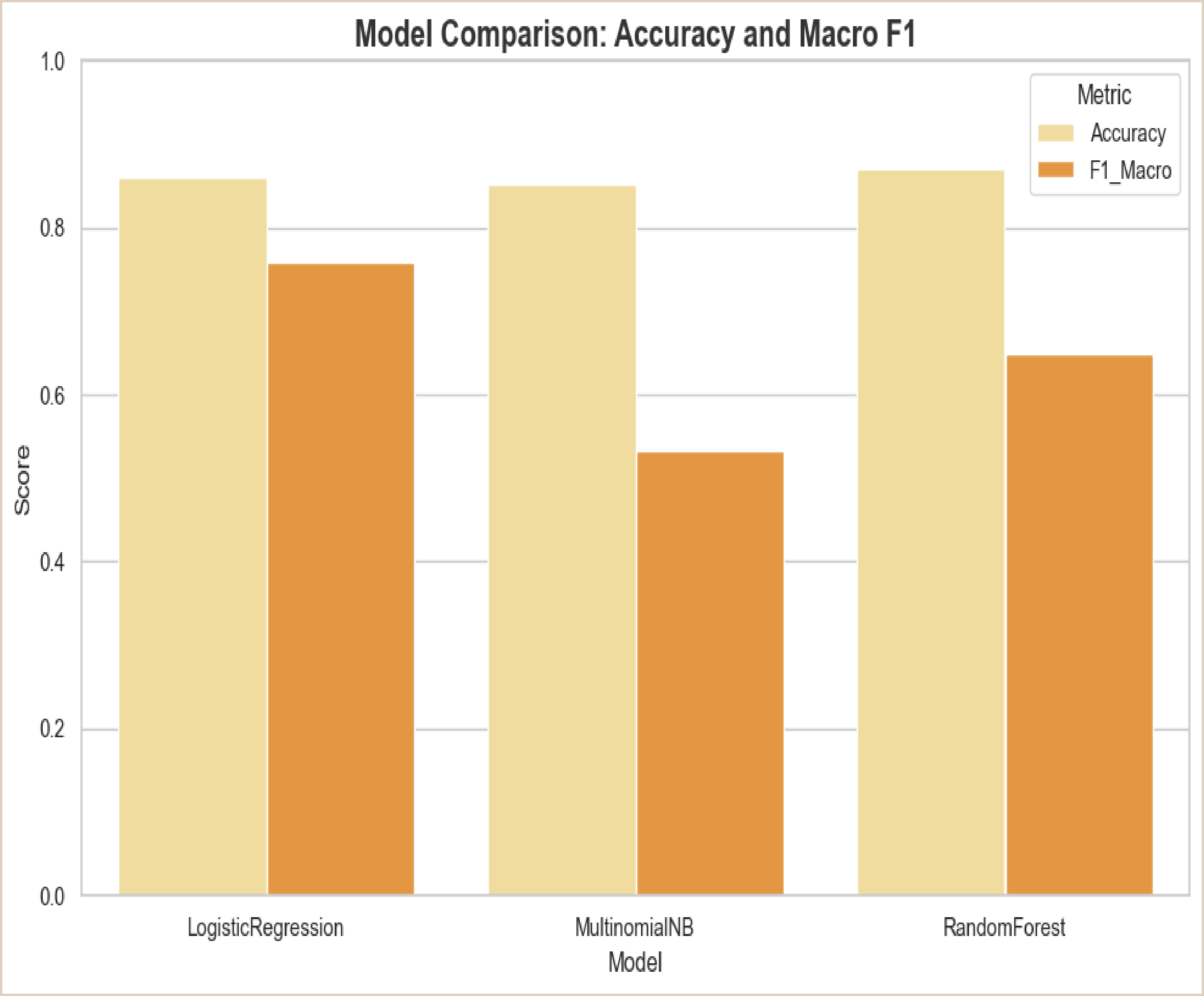
The model classifies tweets into positive and negative categories, excluding neutral and ambiguous data.

Dataset Splitting

Data was split into 80% training and 20% testing, ensuring stratification based on sentiment distribution.

TF-IDF Vectorization

Text data vectorized using TF-IDF with unigrams and bigrams limited to 5,000 features for model input.



Model Training & Evaluation

Logistic Regression Performance

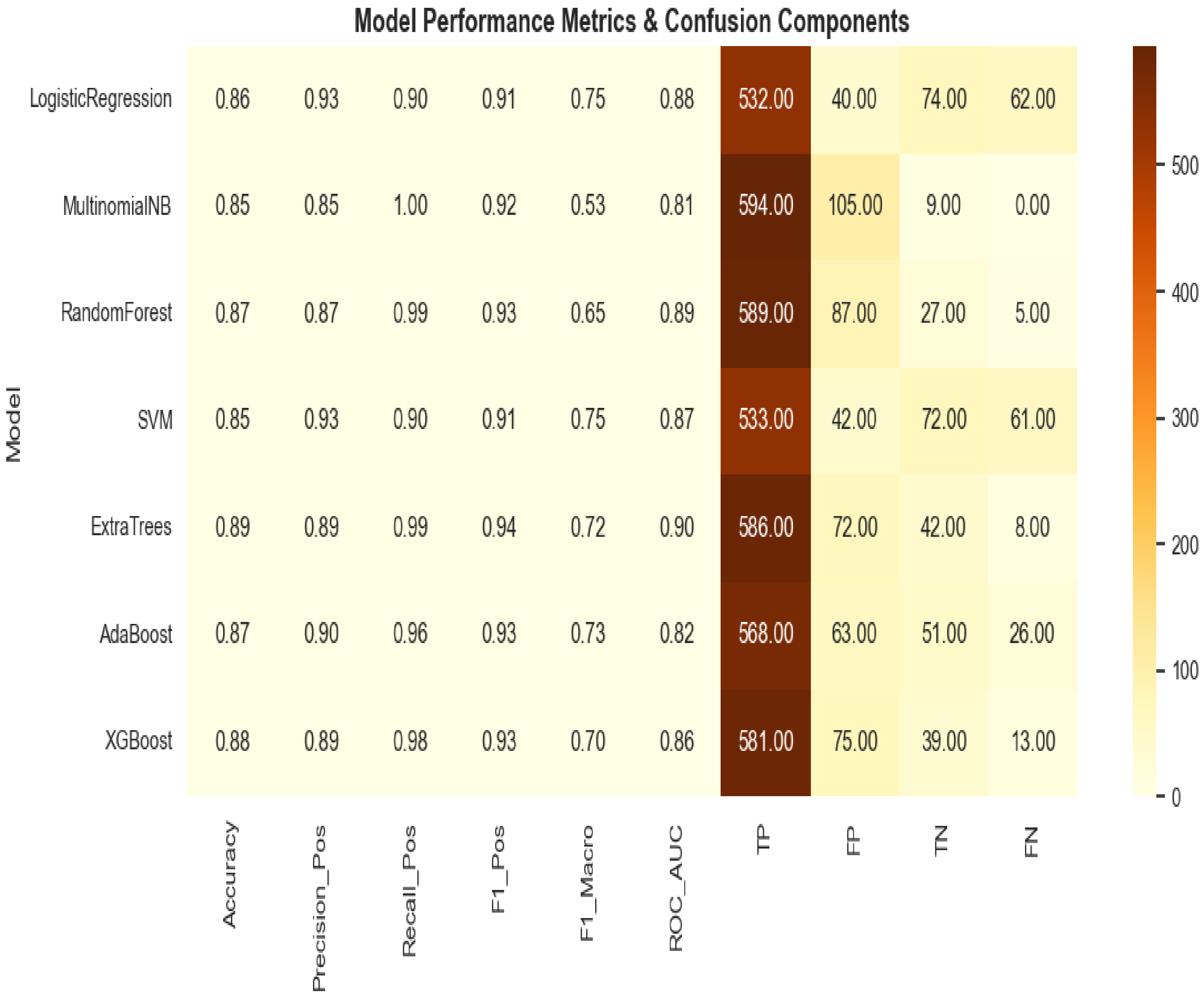
Logistic Regression achieved 86% accuracy and a macro F1 score of 0.76 indicating strong overall model performance.

Naive Bayes Recall

Naive Bayes model demonstrated perfect recall but had a lower macro F1 score of 0.53, indicating imbalance in other metrics.

Random Forest Balance

Random Forest model showed balanced performance with 87% accuracy and a macro F1 score of 0.65.



Model Performance Metrics & Confusion Components

Model	Accuracy	Precision_Pos	Recall_Pos	F1_Pos	F1_Macro	ROC_AUC	TP	FP	TN	FN
LogisticRegression	0.86	0.93	0.90	0.91	0.75	0.88	532.00	40.00	74.00	62.00
MultinomialNB	0.85	0.85	1.00	0.92	0.53	0.81	594.00	105.00	9.00	0.00
RandomForest	0.87	0.87	0.99	0.93	0.65	0.89	589.00	87.00	27.00	5.00
SVM	0.85	0.93	0.90	0.91	0.75	0.87	533.00	42.00	72.00	61.00
ExtraTrees	0.89	0.89	0.99	0.94	0.72	0.90	586.00	72.00	42.00	8.00
AdaBoost	0.87	0.90	0.96	0.93	0.73	0.82	568.00	63.00	51.00	26.00
XGBoost	0.88	0.89	0.98	0.93	0.70	0.86	581.00	75.00	39.00	13.00

- The table shows the performance metrics of different classification models evaluated on a sentiment analysis task. The metrics include accuracy, precision, recall, F1-score, macro F1, ROC AUC, and confusion matrix components.
- **Key Observations**

Best Accuracy

- **Extra Trees (0.89)**** and **XGBoost (0.88)** performed the best in terms of overall accuracy.

Best Precision (Positive class)

- **Logistic Regression (0.93)** achieved the highest precision, meaning fewer false positives.

Best Recall (Positive class)

- **MultinomialNB (1.00)** perfectly captured all positive samples (no false negatives), but at the cost of higher false positives.

Best F1 Score (Positive class):

- **Extra Trees (0.94)** and **(0.93)** gave the best balance between precision and recall.

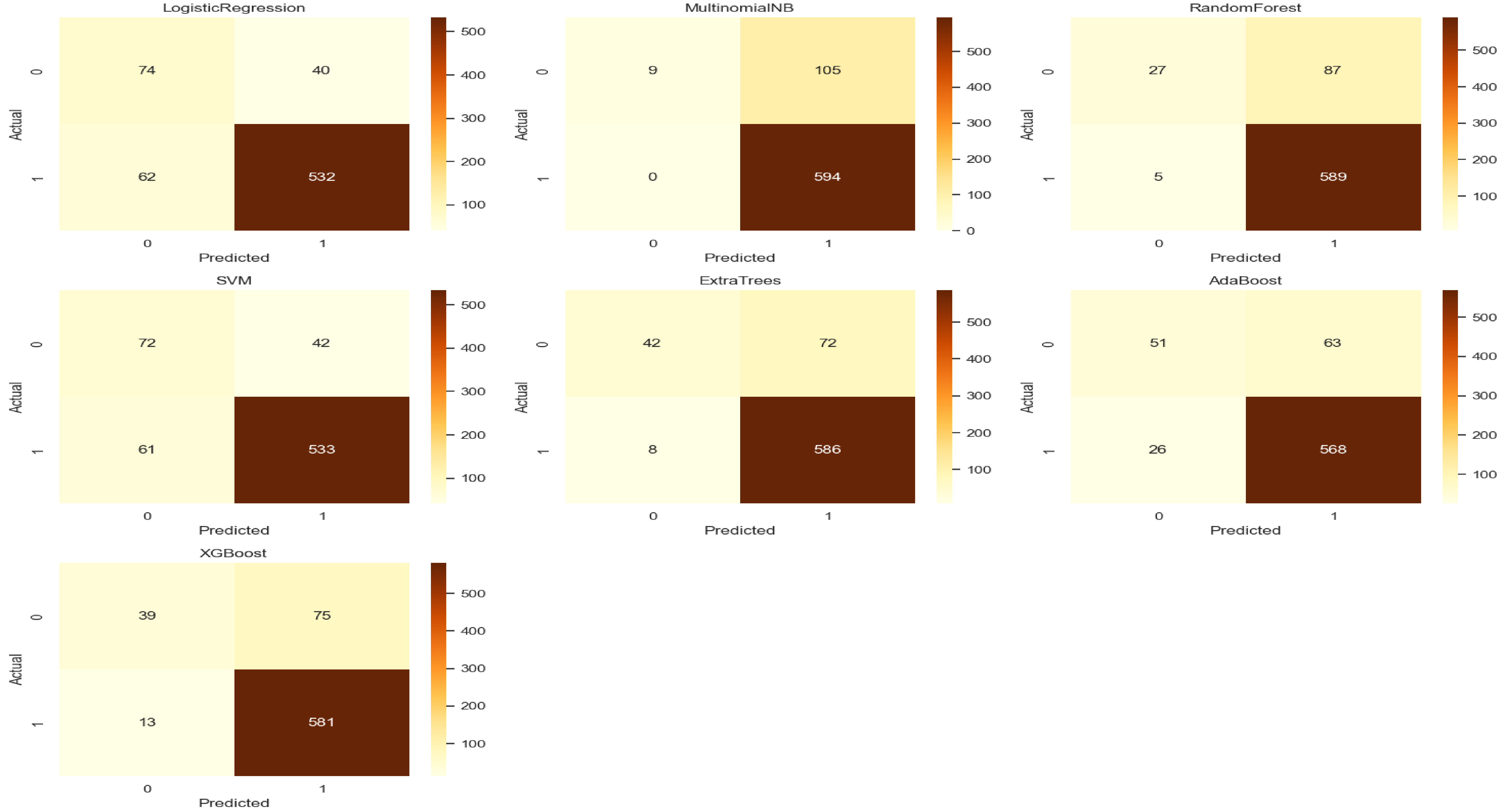
Macro F1:

- **Logistic Regression (0.75)** and **SVM (0.75)** had the best macro F1, meaning they balanced performance across both positive and negative classes better than others.

ROC AUC

- **Extra Trees (0.90)** slightly outperformed others in class separation.

Confusion Matrices for All Models



Confusion Matrices – Summary

Logistic Regression

TN = True Negatives | FP = False Positives | FN = False Negatives | TP = True Positives

- TN=74, FP=40, FN=62, TP=532
- Highest precision (0.93)
- More false negatives

MultinomialNB

- TN=9, FP=105, FN=0, TP=594
- Perfect recall (1.0)
- Very low precision
- **Random Forest**
- - TN=27, FP=87, FN=5, TP=589
- High recall
- Many false positives
- **SVM**
- - TN=72, FP=42, FN=61, TP=533
- Best Macro F1 (0.75, tied)
- Higher false negatives
- **Extra Trees**
- - TN=42, FP=72, FN=8, TP=586
- Best ROC AUC (0.90)
- Strong balance
- Some false positives

AdaBoost

- TN=51, FP=63, FN=26, TP=568
- Balanced trade-off
- Slightly weaker overall

XGBoost

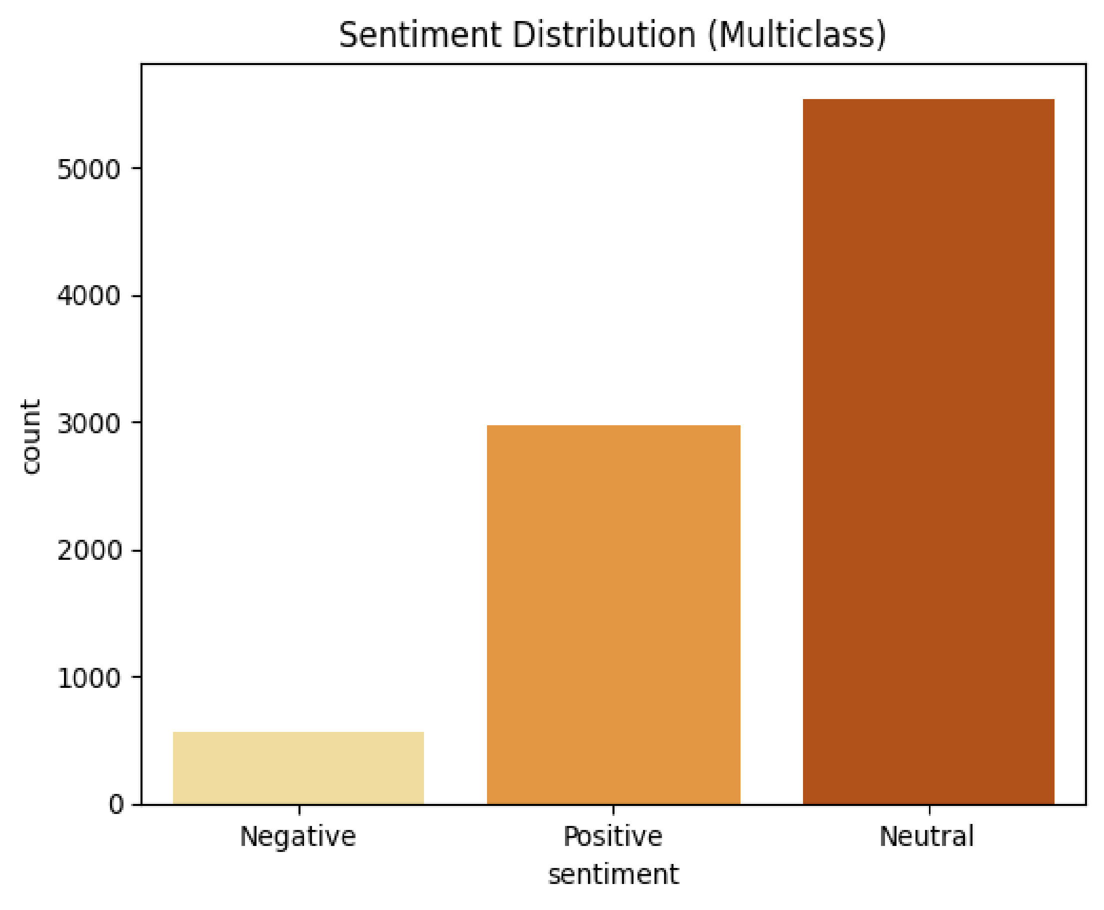
- TN=39, FP=75, FN=13, TP=581
- Strong recall, low FN
- High accuracy & F1
- Some false positives

Highlights

-**Precision** → Logistic Regression (0.93)
- **Recall** → MultinomialNB (1.0)
- **Macro F1** → Logistic Regression & SVM (0.75)
- **ROC AUC** → Extra Trees (0.90)
- **Best Balance** → XGBoost & Extra Trees

Extra Trees and **XGBoost** are strongest overall.
Logistic Regression is safest (few false positives).
MultinomialNB** never misses positives but sacrifices precision.

Multiclass Classification & Training



Multiclass Labeling

Three classes included neutral, positive, and negative tweets with mapped labels and filled missing values.

Data Balancing with SMOTE

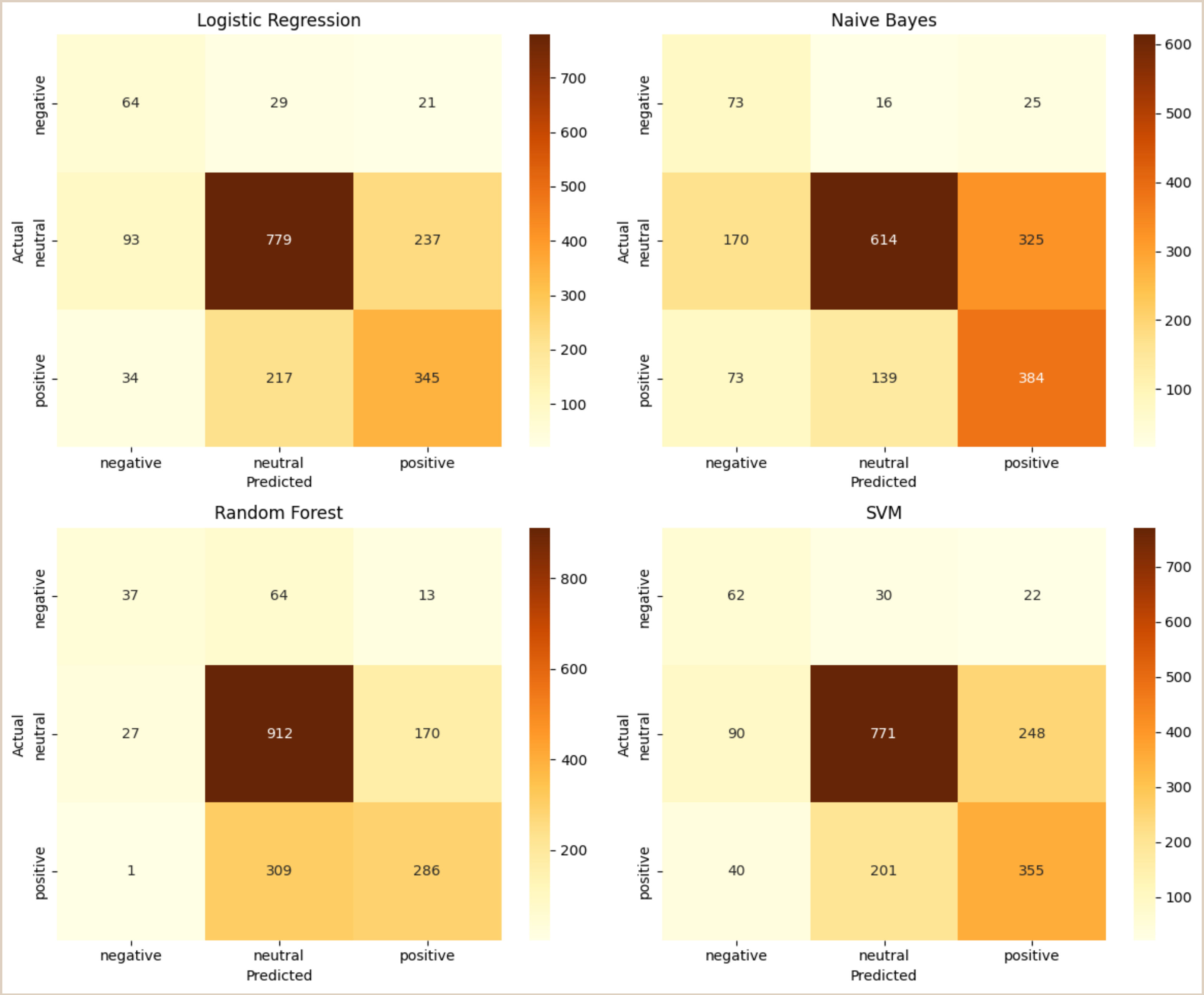
Synthetic Minority Oversampling Technique balanced the dataset to improve model training quality.

Text Preprocessing with TF-IDF

TF-IDF converted text data into numerical features suitable for machine learning models.

Model Training Variety

Models trained included Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machines for classification.



Logistic Regression

	negative	neutral	positive
negative	64	29	21
neutral	93	779	237
positive	34	217	345

Naive Bayes

	negative	neutral	positive
negative	73	16	25
neutral	170	614	325
positive	73	139	384

Random Forest

	negative	neutral	positive
negative	37	64	13
neutral	27	912	170
positive	1	309	286

SVM

	negative	neutral	positive
negative	62	30	22
neutral	90	771	248
positive	40	201	355

Evaluation & Confusion Matrices

Neutral dominates - All models are much better at predicting neutral (majority class).

Precision & recall for neutral are ~0.70–0.82

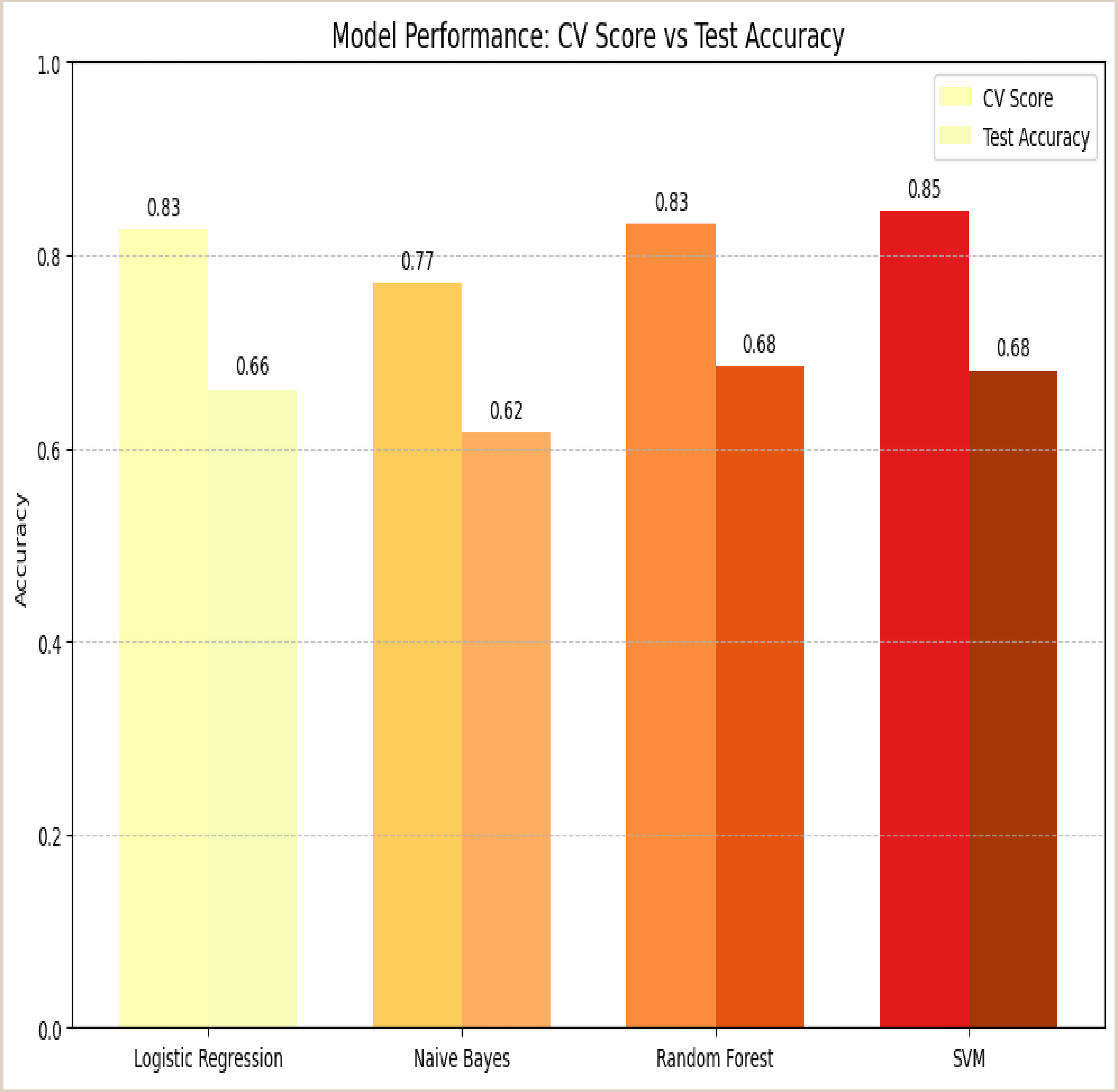
.Minority classes (negative, positive) suffer.

Negative is the weakest class

Logistic Regression: recall 0.56 (ok) but precision only 0.34.

Naive Bayes: recall 0.64, but precision 0.23 (lots of false positives). Random Forest: precision 0.57 but recall drops to 0.32

SVM: similar to Logistic Regression, low precision.



Tuned Model Performance

Hyperparameter Tuning Details

Specific hyperparameters were optimized for each model to improve accuracy and performance.

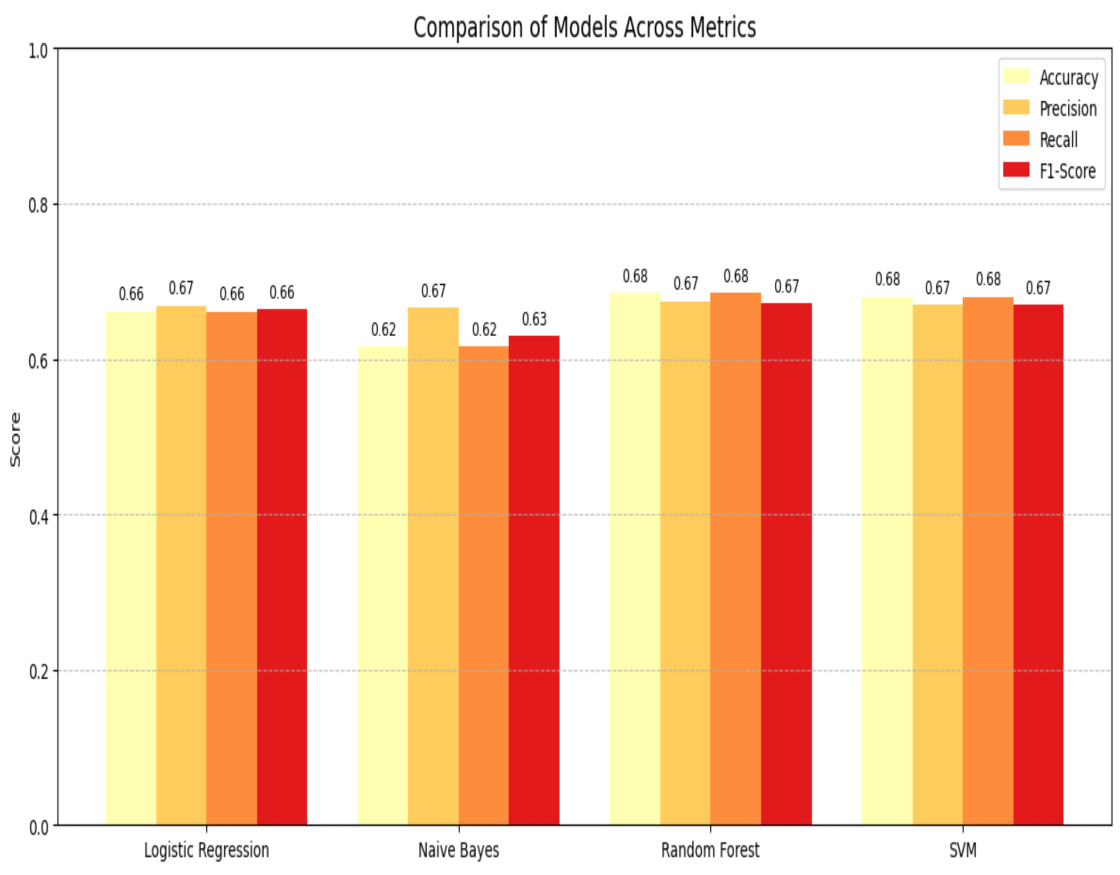
Random Forest Performance

Random Forest achieved the highest test accuracy of 0.684 with balanced evaluation metrics.

Support Vector Machine Results

SVM achieved the highest cross-validation score of 0.846 using kernel 'rbf' and C=8.42.

Final Observations & Best Model



Accuracy: Random Forest and SVM are tied at 0.68 — the highest among all models.

- Precision: Logistic Regression and Naive Bayes are slightly higher (0.67), but overall differences are small.
- Recall: Random Forest and SVM again lead (0.68)
- F1-Score: Random Forest and SVM are slightly higher (0.67), indicating better balance between precision and recall.
- Naive Bayes performs worst overall, likely because it assumes feature independence, which isn't true for your text data.
- Logistic Regression is solid but slightly below Random Forest/SVM in most metrics.
- Random Forest vs SVM: Both have similar scores across all metrics. Random Forest is usually faster to train, easier to interpret (feature importance), and less sensitive to scaling of features. SVM can be slower on large datasets, and tuning kernel parameters can be tricky.

Best model:

- Random Forest is the best model because: - It has top test accuracy (0.68). - High recall and balanced F1-score. - More robust and interpretable than SVM. - Handles feature correlations better than Naive Bayes and logistic regression.

Recommendation

Implement the Random Forest model for sentiment classification of tweets. It demonstrated the best balance of accuracy (68%), precision, recall, and interpretability among all tested models. Its robustness and ability to handle feature correlations make it ideal for real-world deployment in monitoring brand sentiment.

QUESTIONS