SAFARICOM/AIRTEL APP SENTIMENT ANALYSIS

Presented by:
Anthony, Winnie, Jedidah,
Navros, Rachel, Anne.

PROJECT OVERVIEW

- This project leverages Natural Language Processing (NLP) to analyze customer sentiment in Safaricom and Airtel apps (M-PESA, MySafaricom, Airtel Money, MyAirtel).

- We developed predictive models to classify sentiment into Positive, Neutral, or Negative.

- The goal is to deliver a proof-of-concept system that not only predicts sentiment but also provides actionable insights for telco decision-making.

STAKEHOLDER AND BUSINESS VALUE

Key Stakeholders

- Safaricom & Airtel Product Teams – M-PESA, Airtel Money, MySafaricom, MyAirtel
- Customer Experience & Support Teams – act on user complaints and feedback
- Marketing & Brand Managers – monitor brand sentiment and campaigns
- Senior Executives & Decision-Makers – guide strategy and investments
- Data Science & Analytics Teams – maintain models, monitor trends, and deliver insights

Business Value

- Customer Insights: Reveal what drives satisfaction vs. frustration.
- Faster Response: Detect negative feedback early and act before issues escalate.
- Retention & Revenue: Reduce churn by resolving customer pain points quickly.
- Strategic Decisions: Guide feature updates, bundle pricing, and app improvements.
- Continuous Learning: Empower data scientists to retrain models and track sentiment shifts over time.

PROJECT PIVOT

Challenge

- We initially planned to analyze tweets about Safaricom and Airtel.
- Twitter API limits restricted access to enough reliable data.

Pivot

- Shifted to Google Play & App Store reviews for M-PESA, MySafaricom, Airtel Money, and MyAirtel.

Advantage

- App-specific: Direct link to real user experience.
- Richer: More detailed than short tweets.
- Ongoing: Reviews update continuously for easy monitoring.

👉 Result: A cleaner, richer dataset aligned with our business goals.

ABOUT THE DATASET

Source

- Google Play & App Store reviews
- Apps Covered:
- Safaricom → M-PESA, MySafaricom
- Airtel → Airtel Money, MyAirtel

Dataset Overview

- Size: 5,000+ reviews
- Scope: Ratings (1–5★), text reviews, likes, platform, date

Main Columns

- date, app_name, platform, rating, review, sentiment

Feature-Engineered Columns

- review_length, word_count, clean_text,

LIMITATIONS OF THE DATASET

- More reviews on Google Play than App Store , slight platform bias.
- Extreme opinions dominate (very happy or unhappy users).
- Short reviews reduce text analysis depth.
- Mixed languages (English, Swahili, Sheng) affect accuracy.
- No user demographics, can’t segment by age or region.

👉 Despite these limits, cleaning and balancing kept results reliable.

DATA CLEANING AND PRE- PROCESSING

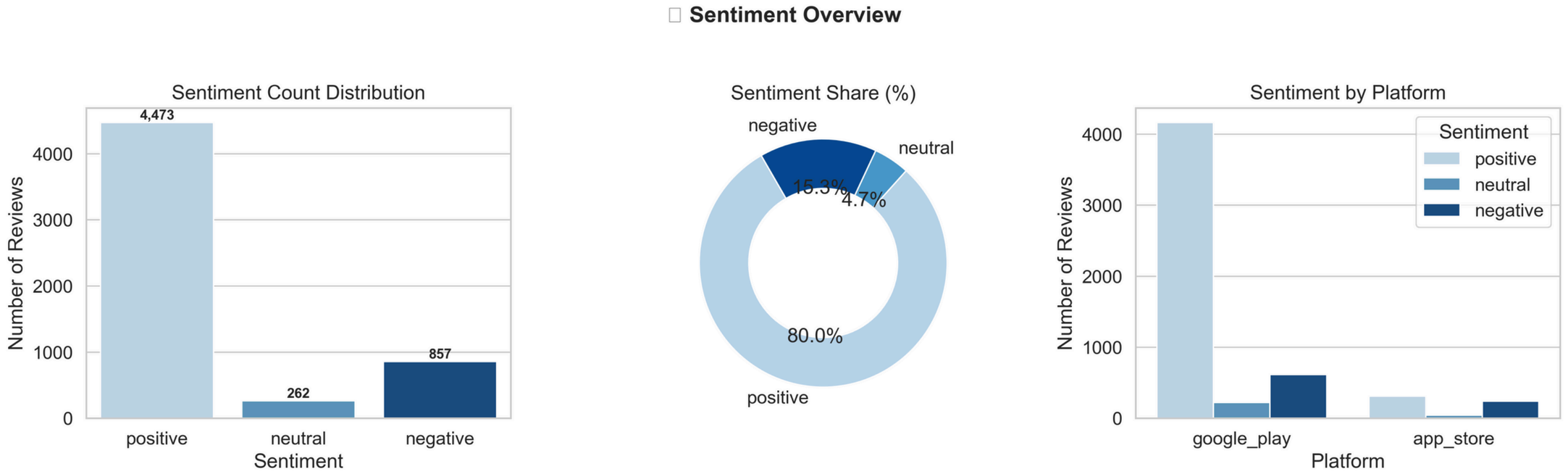
- Removed missing & duplicate data → ensured accuracy and relevance
- Handled missing brand references → assigned “Unknown” to keep all records
- Standardized text → lowercased, removed punctuation, URLs, hashtags, mentions, and numbers

- 🔧 Preprocessing for NLP:
- Tokenization, stopword removal, and lemmatization
- TF-IDF vectorization (unigrams & bigrams)
- Checked class balance → negatives underrepresented → applied SMOTE for binary tasks
- Encoded labels (Positive, Neutral, Negative)

Outcome:

- Ensured data quality and reliable input for feature engineering & model training

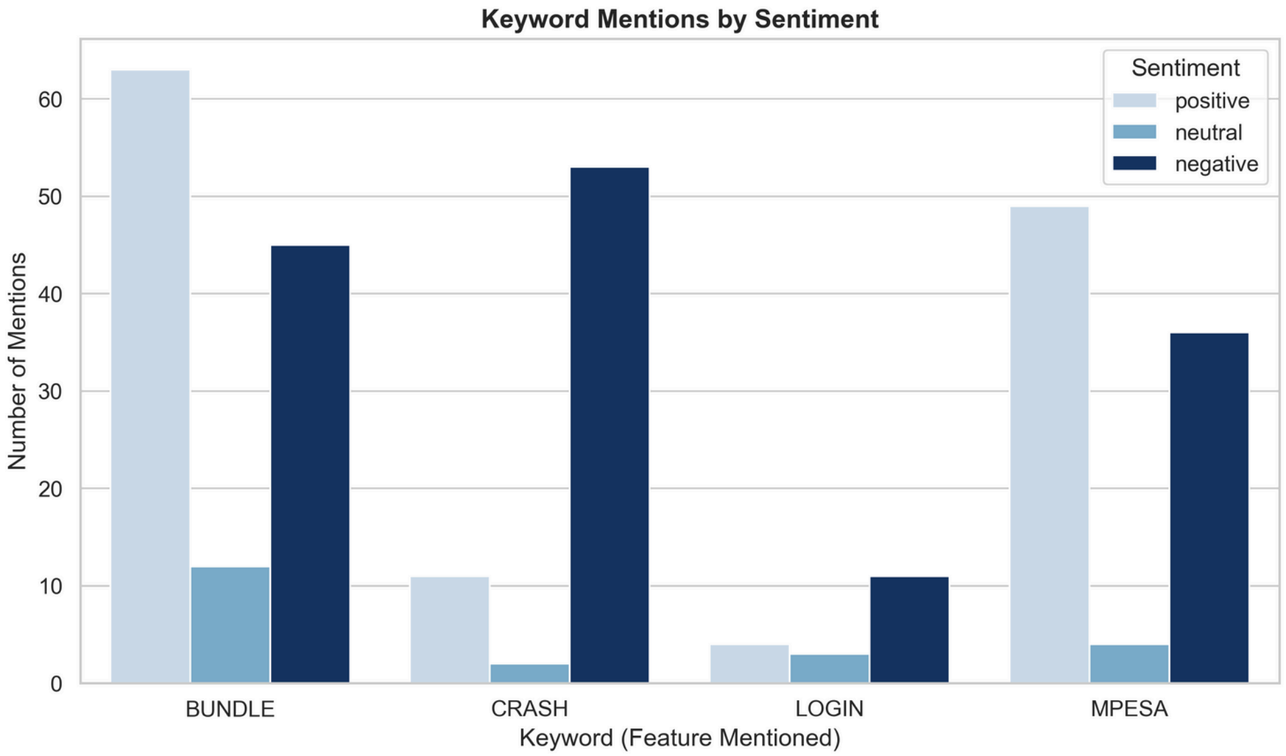
KEY INSIGHTS FROM EDA

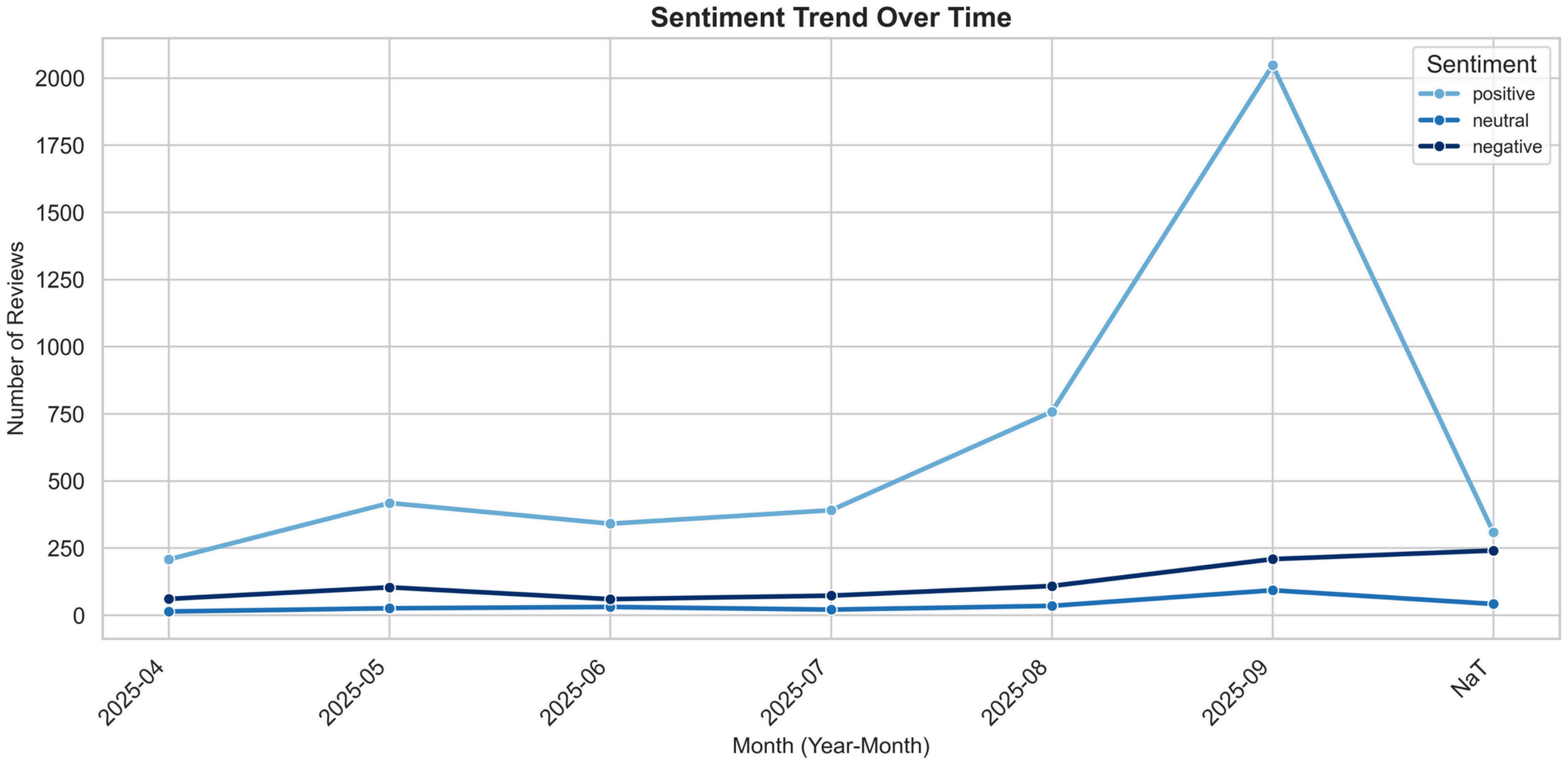


- Overall: Majority of reviews are Positive (~80%)
- Neutral: Small proportion, often ambiguous or mixed tone
- Negative (~15%) highlight recurring pain points (crashes, bundles, login, billing)
- Insight: While sentiment skews positive, negative voices carry higher impact

KEY INSIGHTS FROM EDA

- Bundle: Most frequently mentioned; praised for convenience but criticized for performance & pricing
- Crash: Strongly tied to negative reviews, confirming stability issues
- Login: Appears less often but mostly negative, highlighting authentication frustrations
- M-PESA: Dominantly positive, reinforcing reliability & trust in payments



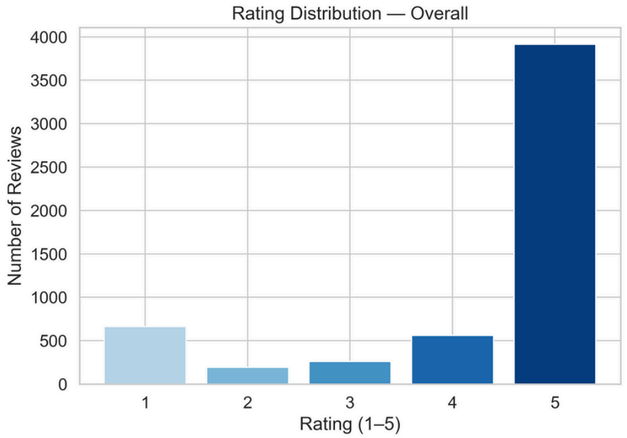


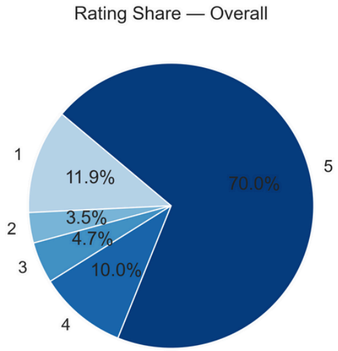
- Positive reviews dominate consistently across months
- Dips in sentiment observed after some app updates → possible bugs or rollout issues
- Negative reviews spike during high-traffic periods, showing frustrations spread fast
- Neutral remains stable, reflecting mixed but less vocal users

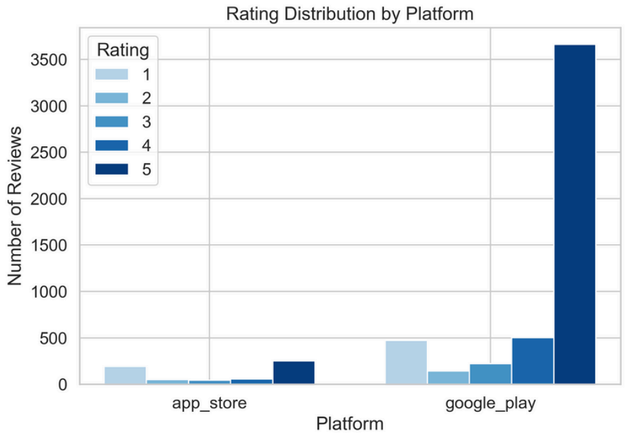
KEY INSIGHTS FROM EDA

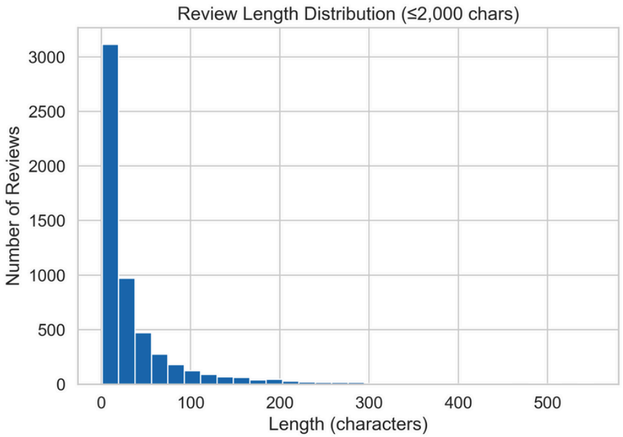
- Most reviews are very positive:
  - 70% gave a 5-star rating.
  - Only 12% gave a 1-star rating.
- Platform differences:
  - Google Play dominates in both number of reviews and share of 5-star ratings.
  - App Store has far fewer reviews overall.
- Review length:
  - Majority of reviews are short (under 50 characters).
  - Shows that users often leave quick feedback rather than detailed text.

👉 Conclusion:
Overall ratings are highly positive, especially on Google Play, but negative reviews (1-star) still represent a meaningful group to pay attention to.









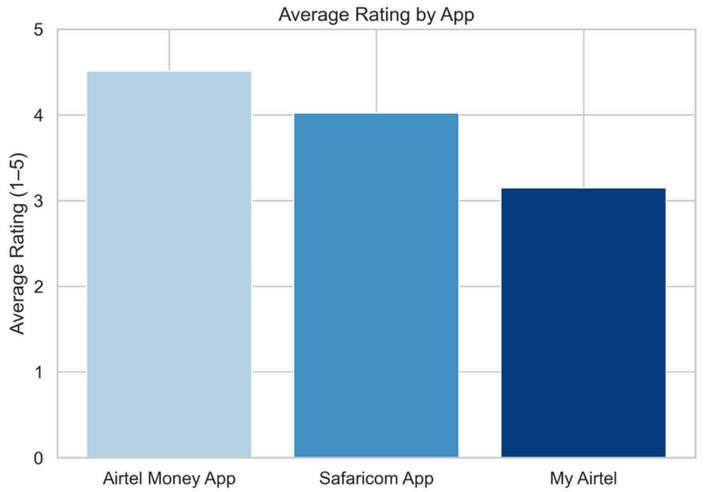
KEY INSIGHTS FROM EDA

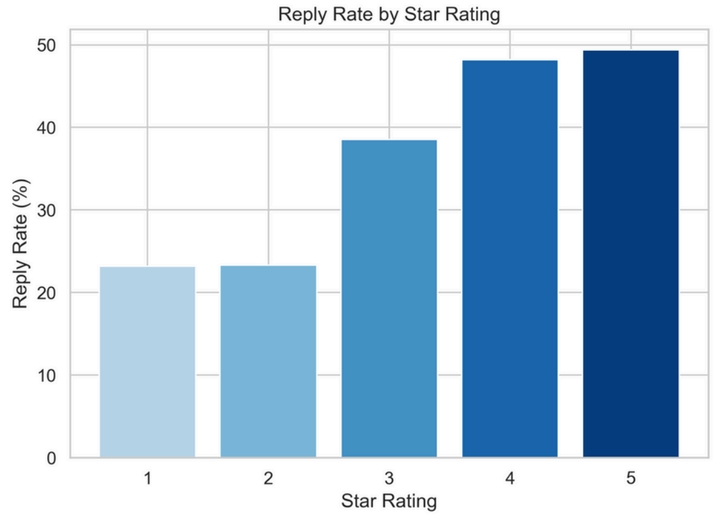
App Ratings:
- Airtel Money leads (~4.5★), Safaricom solid (~4.0★), MyAirtel lags (~3.1★)

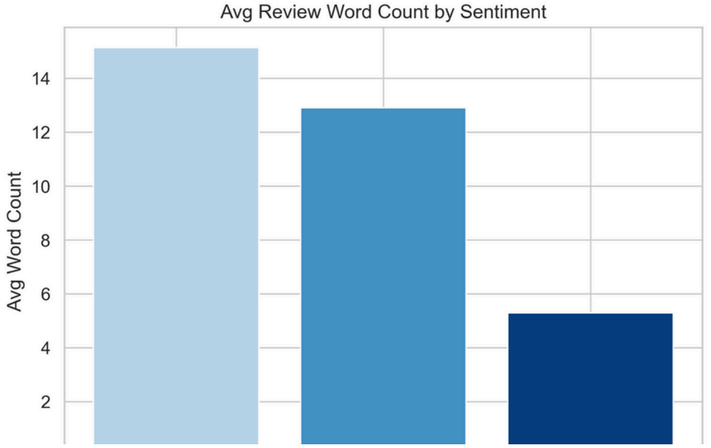
Reply Rates:
- Higher for positive reviews (~50%) vs low for negatives (~23%) → missed recovery opportunity

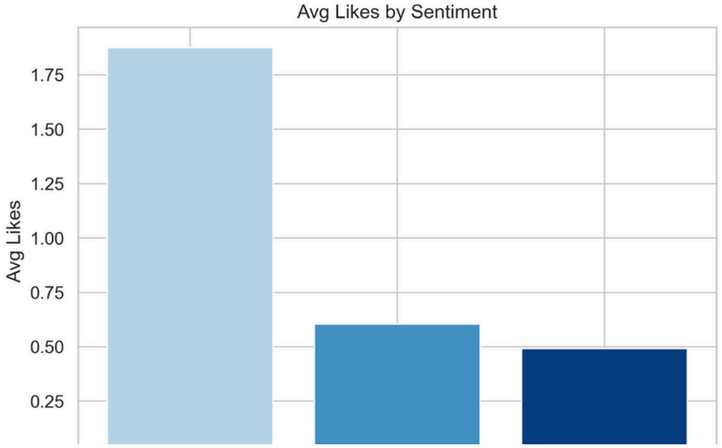
Review Length:
- Negative reviews are longest → detailed complaints
- Positive reviews short → generic praise

Likes:
- Negative reviews get most likes → frustrations resonate widely









Word Clouds — Positive vs Negative Reviews

Positive Reviews Word Cloud

Negative Reviews Word Cloud

Positive Reviews:

- Dominated by words like “good,” “best,” “awesome,” “excellent,” “easy,” “mpesa”
- Users appreciate speed, reliability, and convenience

Negative Reviews:

- Frequent terms include “open,” “work,” “service,” “data,” “update,” “bundle”
- Indicates frustration with network reliability, app crashes, and updates

MODELING APPROACH AND PIPELINE

Modeling Goals

- Build models to predict customer sentiment
- Start with Binary Classification → Positive vs. Negative
- Extend to Multiclass Classification → Positive, Neutral, Negative
- Compare traditional ML vs advanced NLP (BERT)
- Deploy a real-time sentiment prediction API

Pipeline Overview

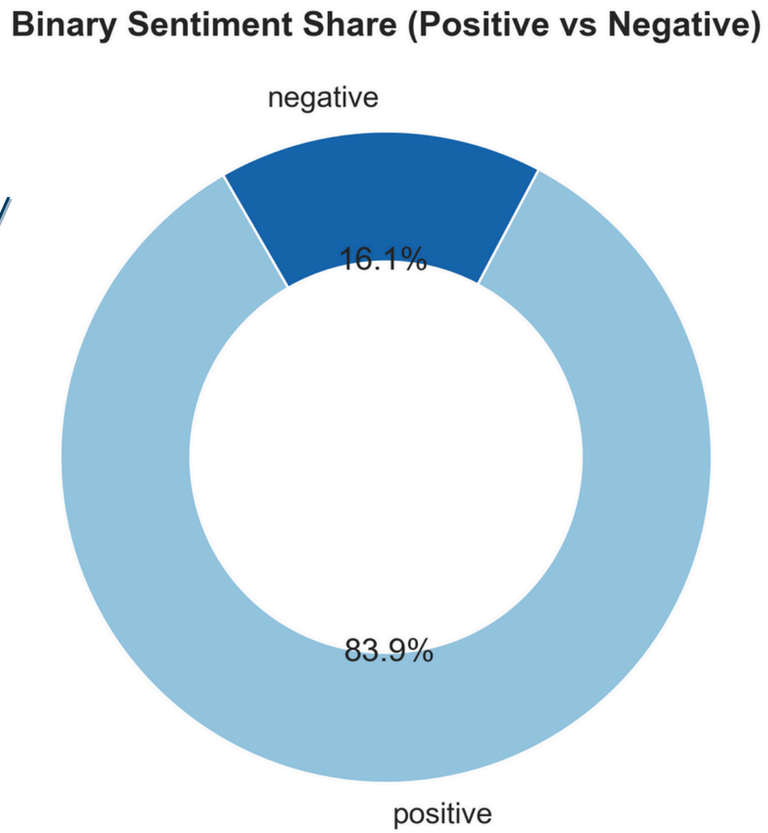
1. Text Preprocessing: Tokenization, stopword removal, lemmatization
2. Feature Extraction: TF-IDF (unigrams & bigrams)
3. Model Training:

1. Balancing: SMOTE applied for underrepresented Negative class
2. Evaluation: Accuracy, F1-score, ROC-AUC, Confusion Matrices
3. Deployment: FastAPI + Render for live predictions

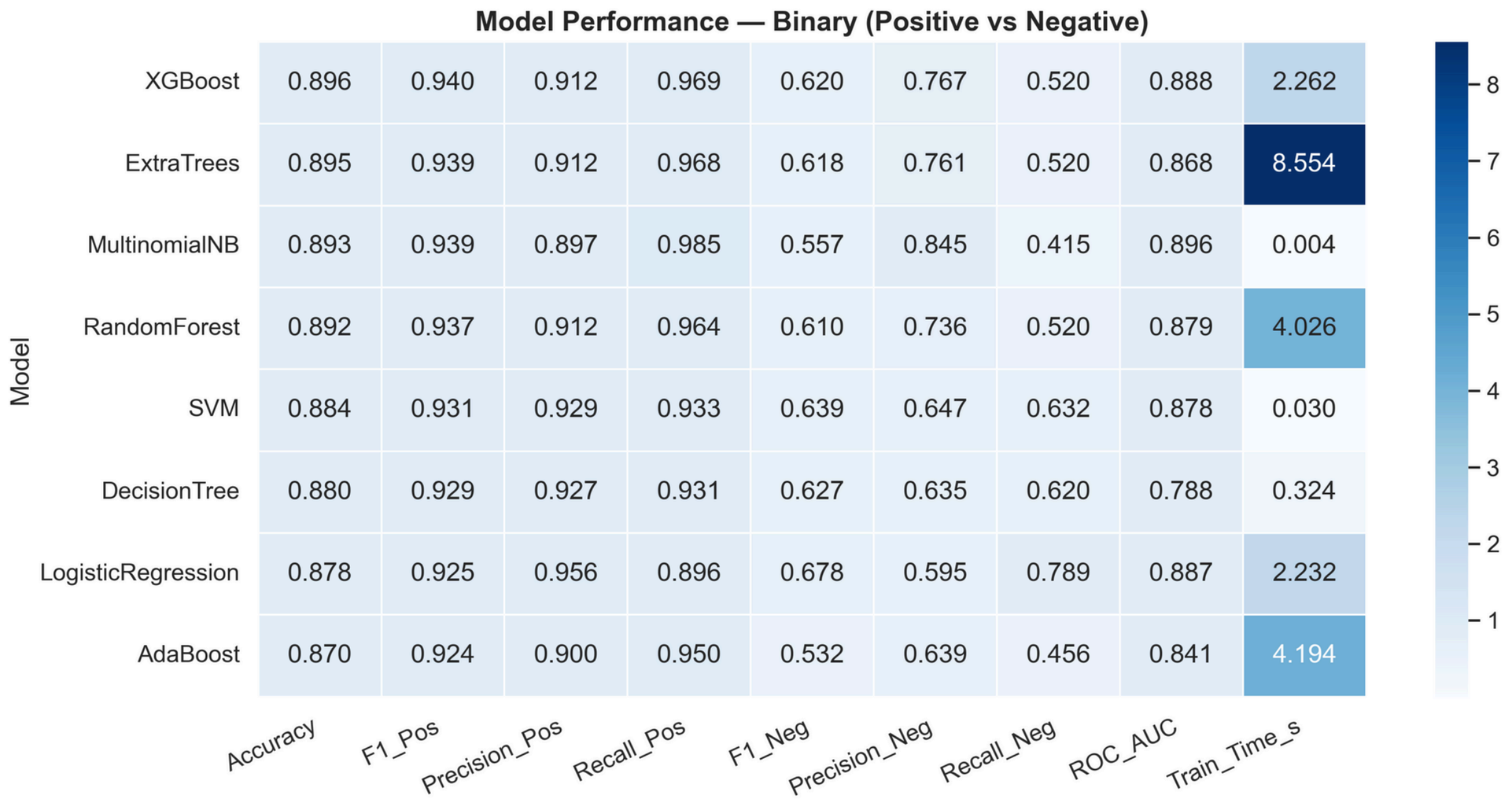
BINARY CLASSIFICATION

Binary Sentiment Modeling (Positive vs Negative)

- Objective: Build a robust baseline model for customer polarity
- Challenges:
  - Highly imbalanced data (Positive ≫ Negative)
  - Short and noisy review texts
- Approach:
  - Applied SMOTE and class weighting to balance labels
  - Trained multiple models → Logistic Regression, SVM, Random Forest, AdaBoost, ExtraTrees, DecisionTrees, MultinomialNB, XGBoost
  - Evaluated via Accuracy, Precision, Recall, F1-score, and ROC-AUC
- Result:
  - Logistic Regression (with SMOTE + weighting) gave the best balance of performance and interpretability



MODEL PERFORMANCE

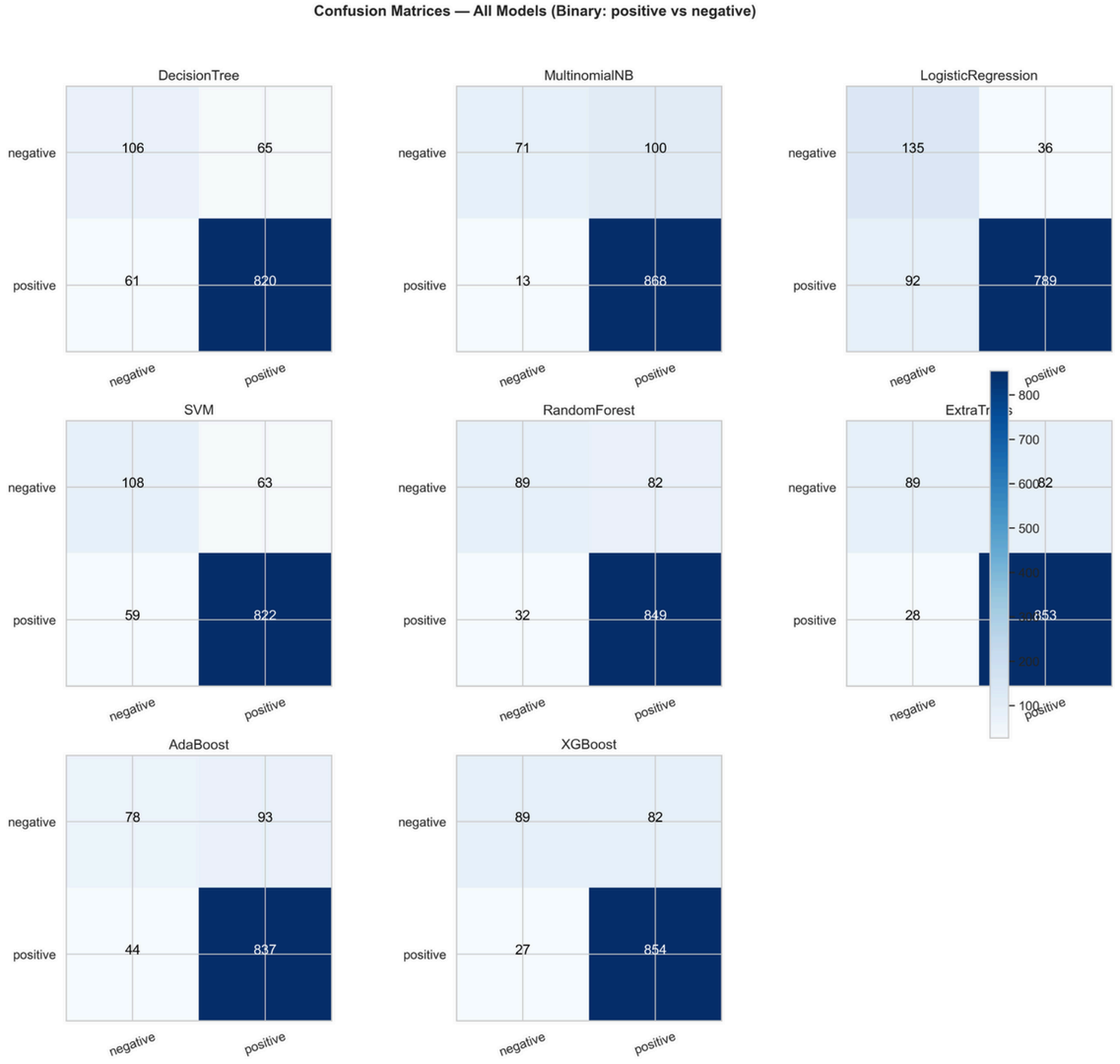


MODEL HIGHLIGHTS

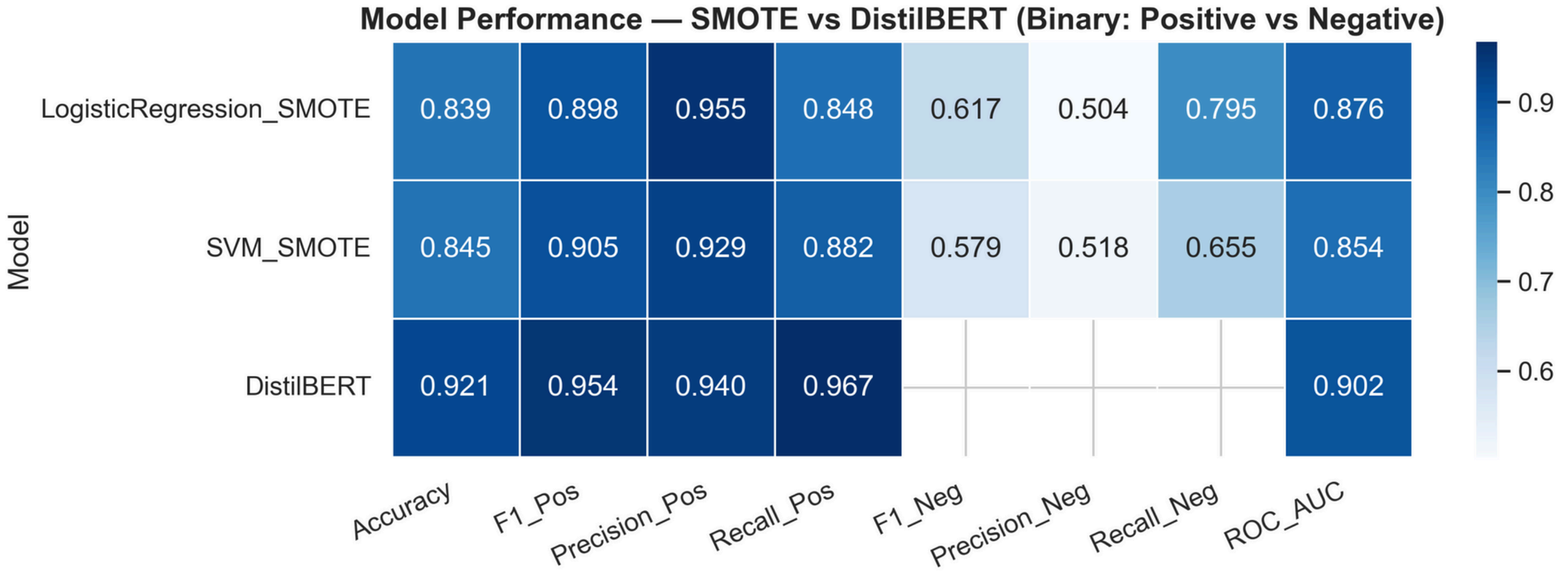
- All models did well – over 87% accuracy.
- XGBoost and ExtraTrees are the most reliable overall, correctly classifying almost 9 out of 10 reviews.
- For positive reviews: XGBoost and ExtraTrees are best at finding them. Logistic Regression is the most precise (fewest false alarms).
- For negative reviews (complaints): SVM and Decision Tree are best at spotting more of them, while Naive Bayes is good at being precise but misses many.
- Overall balance: XGBoost and Logistic Regression separate positives and negatives most clearly.
- Training speed: Naive Bayes and SVM are extremely fast; XGBoost is still fast enough while giving the best all-round results.

CONFUSION MATRICES

- Dark boxes = correct predictions; light boxes = model confusion.
- Positive reviews dominate — all models detect them well.
- Negatives are harder to catch:
- XGBoost/ExtraTrees: Strong overall, miss some negatives.
- SVM/Decision Tree: Better at spotting complaints.
- Naive Bayes: Precise but misses many negatives.
- Logistic Regression: Balanced and easy to explain.
- 👉 Conclusion:
- XGBoost = best overall, SVM/Decision Tree = best for complaints, Naive Bayes = fastest, Logistic Regression = reliable baseline.



SMOTE AND BERT MODELS



- Logistic Regression (SMOTE): Good on positives, but misses some complaints.
- SVM (SMOTE): Slightly better, but still struggles with complaints.
- DistilBERT: Best performer – 92%+ accuracy, strong on both positives and negatives.
- 👉 Conclusion:
- DistilBERT = most accurate.
- SMOTE models = lighter, faster, easier to deploy.

MODEL OF CHOICE

- Balanced Trade-off: While DistilBERT gives the highest accuracy, it requires heavy resources and longer run-times. Logistic Regression with SMOTE offers a good balance of performance and efficiency.
- Handles Class Imbalance: SMOTE boosts performance on the minority class (complaints), so we don't miss as many negative reviews.
- Fast & Lightweight: Runs quickly, easy to retrain, and efficient to deploy on available infrastructure.
- Interpretable: Unlike complex models, Logistic Regression is simple to explain to stakeholders, regulators, and business users.
- Strong Results: Delivers solid accuracy (>83%) and good separation of positive vs negative sentiment.
- 👉 Conclusion: Logistic Regression with SMOTE was selected as the best practical model — it balances accuracy, speed, interpretability, and ease of deployment.

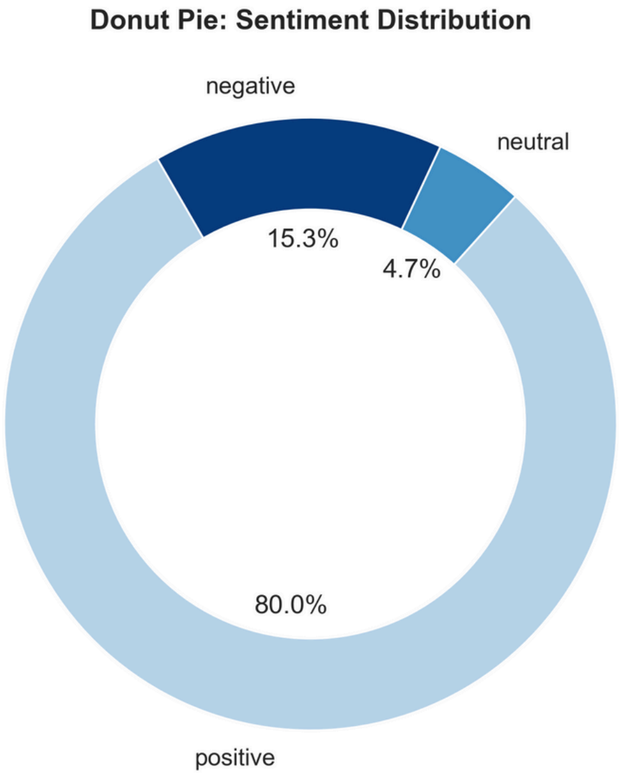
MULTI-CLASS CLASSIFICATION

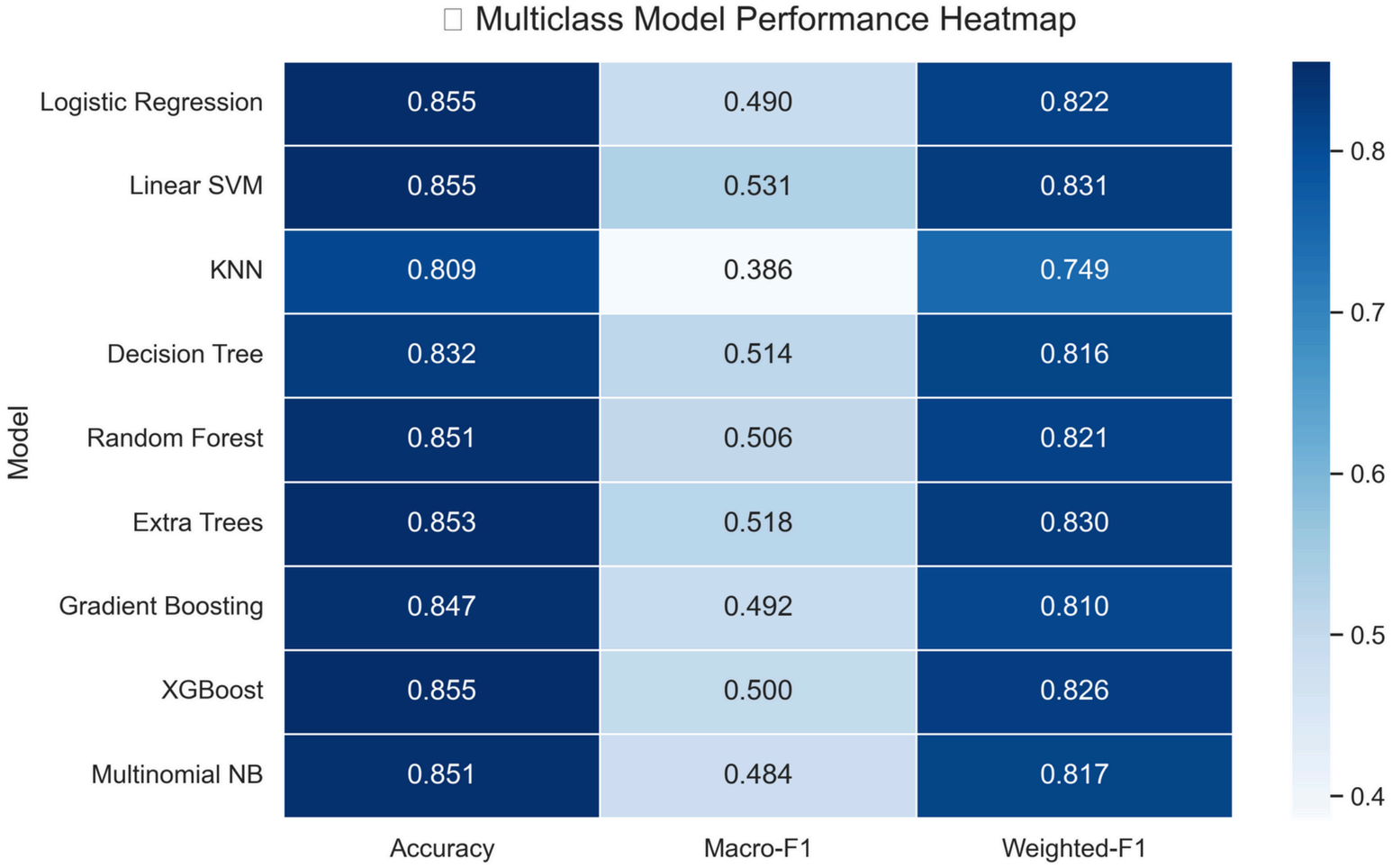
We transitioned from a Binary (Positive/Negative) classification to a Multiclass (Positive, Neutral, Negative) classification to capture more feedback.

- Neutral reviews help reveal mixed opinions often missed before.
- This gives a fuller picture of customer emotions and helps identify areas for improvement.

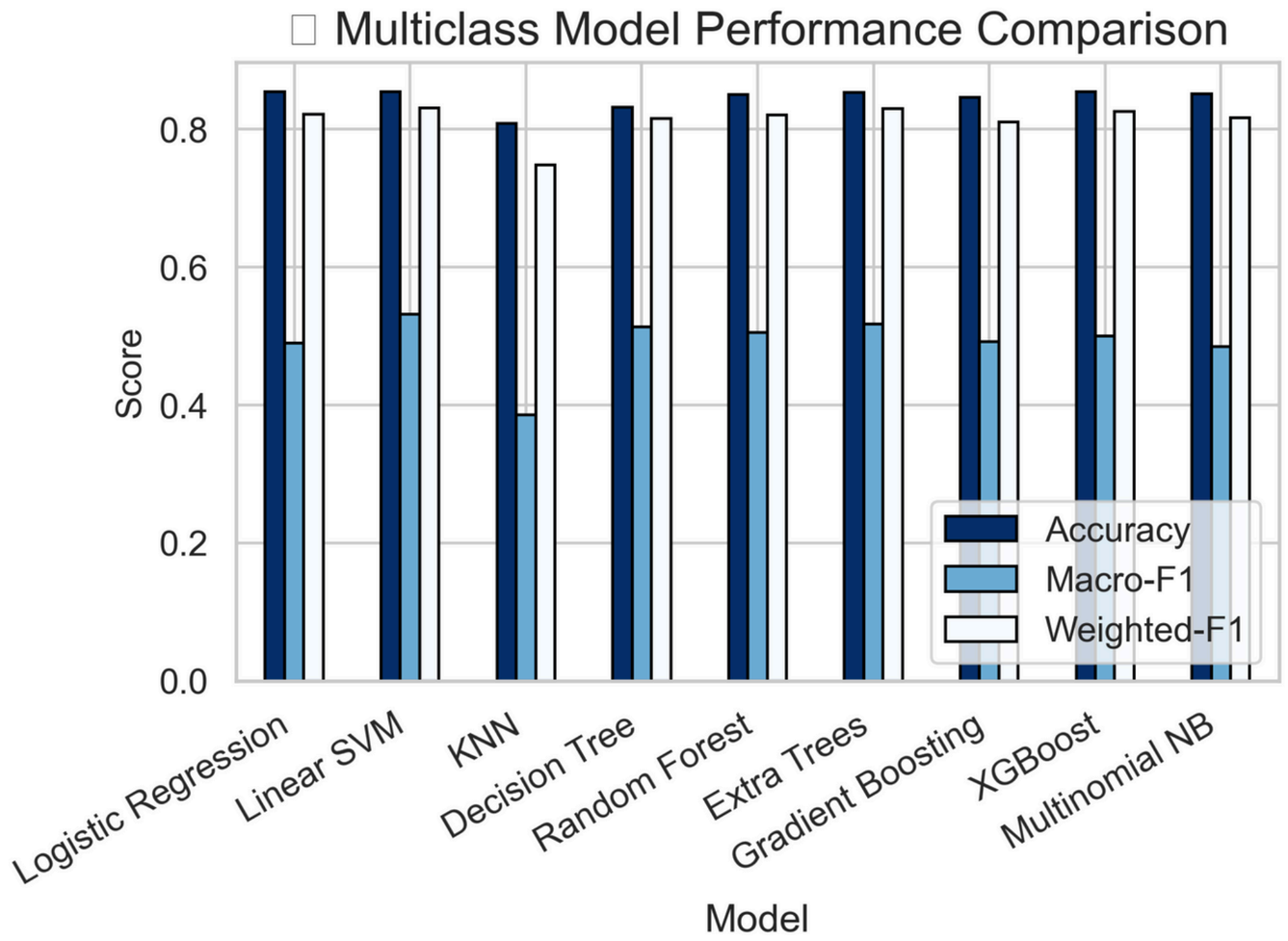
👉 Goal:

- Get a balanced, 3-way understanding of customer sentiment, not just good or bad.

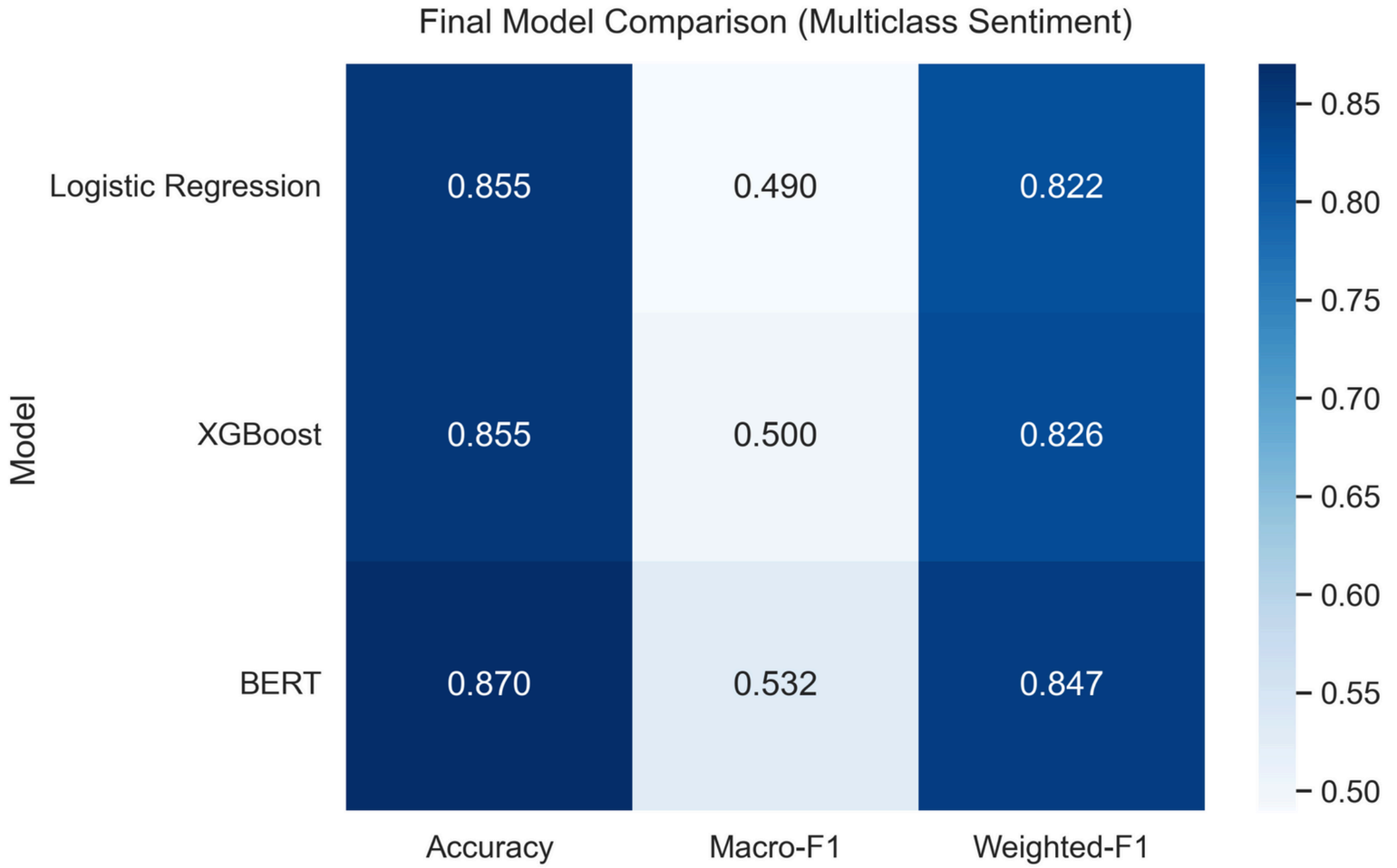




- Overall accuracy across models ranged from 80%–86%, showing consistent performance.
- XGBoost, ExtraTrees, and Logistic Regression performed best with ~85% accuracy.
- Macro-F1 scores (0.49–0.53) show the challenge of correctly classifying neutral reviews, which often overlap with positive or negative tones.
- Weighted-F1 (0.82–0.83) indicates models handled class imbalance well overall.
- 👉 Summary:
- Models are stable and reliable across three sentiment classes, with XGBoost and Logistic Regression emerging as top consistent performers.
-



- Model Comparison (Accuracy, Precision, Recall, F1)
- Random Forest & SVM achieve highest scores (~0.66 across metrics).
- Logistic Regression: Stable mid-range performer (~0.63–0.65).
- Naive Bayes: Lowest performance (F1 ≈ 0.59).
- Best balance: Logistic Regression & SVM.



- BERT performs the best overall, capturing subtle tone differences, especially in neutral reviews.
- XGBoost and Logistic Regression are nearly tied in performance, with only slight differences in F1.
- BERT = highest performance, but requires more computation and training time.
- Logistic Regression = simpler, faster, and interpretable, making it ideal for deployment.

MODEL DEPLOYMENT

- Deployed Model: Logistic Regression (with SMOTE)
- Why:
  - High accuracy and balanced predictions.
  - Lightweight, fast to run, and easy to integrate.
  - SMOTE helped improve detection of minority (negative) reviews.
  - Clear coefficients for interpreting which words drive each sentiment.

🔹 Deployment Setup

- Backend: FastAPI (for real-time prediction)
- Hosting: Render / Hugging Face Space (scalable & accessible)
- Input: User text review
- Output: Predicted Sentiment → Positive, Neutral, or Negative

👉 Business Benefit: Enables instant sentiment classification for new customer feedback, supporting real-time monitoring of brand perception.

BUSINESS IMPACT

Customer Insights

- Sentiment is mostly positive , users are satisfied.
- Neutral and negative reviews show where to improve.
- Neutral users are easy wins for conversion.

Model Insights

- Models hit 85%+ accuracy.
- BERT is most precise; Logistic Regression (SMOTE) is fastest and best for live use.
- Enables early issue detection with speed.
-

Business Value

- Real-time sentiment tracking across platforms.
- Quick response to customer pain points.
- Actionable insights for product, marketing, and support.

👉 In short:
We turned reviews into real-time business intelligence.

CONCLUSION

Immediate Actions

- Deploy model to auto-classify new customer feedback in real time.
- Monitor positive & negative reviews for quick follow-up.
- Launch a simple dashboard for live sentiment tracking.

Next Phase

- Expand analysis to other regions & languages.
- Upgrade to DistilBERT/BERT once infrastructure grows.
- Schedule monthly retraining to keep predictions current.

👉 Takeaway:

We’re turning sentiment analysis into a strategic business tool, powering faster decisions, happier customers, and smarter growth.

"EVERY DATASET HAS A STORY — AND TODAY, WE'VE JUST TOLD OURS."

Q N A

THANK YOU

PROUDLY PRESENTED BY GROUP FIVE — WHERE DATA MEETS IMPACT.