Twitter Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Understanding Brand Perception Through Machine Learning & NLP

Group 1

Muema Stephen · Salma Mwende · Grace Wangui

Machine Learning · Sentiment Classification · Social Media Analytics

Table of Contents

Business Problem & Context

The Challenge

In today's digital marketplace, social media platforms like Twitter serve as primary channels where customers express opinions about products and brands. Companies face critical challenges:

- **Volume Overload:** Thousands of brand mentions daily make manual analysis impossible
- **Response Delays:** Traditional monitoring identifies issues only after brand damage occurs
- **Resource Constraints:** Customer service teams overwhelmed by feedback volume
- **Prioritization Issues:** Difficulty distinguishing urgent complaints from general chatter

Our Solution

Build an automated sentiment classification system that processes Twitter data in real-time, categorizing tweets into three sentiment classes:

Positive	Negative	Neutral
Enthusiastic endorsements	Complaints & issues	Informational mentions

Business Value

Real-Time Crisis Detection
Identify emerging negative sentiment before it escalates

Automated Monitoring
Scale from hundreds to thousands of tweets daily

Smart Prioritization
Route critical complaints to specialized response teams

Data-Driven Insights
Measure campaign effectiveness quantitatively

Success Metrics

We aim to build a model that can accurately classify tweet sentiment with high precision and recall, particularly for **negative sentiment** where misclassification is most costly from a business perspective.

Project Objectives & Success Metrics

Main Objective

Develop an end-to-end Natural Language Processing pipeline for automated multi-class sentiment classification that accurately categorizes Twitter data into **Positive**, **Negative**, and **Neutral** sentiment classes, enabling real-time brand monitoring and crisis detection.

Specific Goals

Handle Class Imbalance
Implement techniques to address the 61%-33%-6% distribution

Robust Negative Detection
Maximize recall for negative sentiment (critical for crisis detection)

Real-Time Processing
Enable fast inference suitable for streaming Twitter data

Actionable Insights
Provide business-relevant sentiment analysis and recommendations

Success Criteria

Overall Accuracy — Target: >65%
Correct classification rate across all sentiments

Negative Recall — Target: >45%
Percentage of actual negatives correctly identified

Macro F1-Score — Target: >0.55
Balanced performance across all sentiment classes

Cross-Validation Stability — Low Variance
Consistent performance across different data splits

Dataset Overview & Characteristics

Source & Context

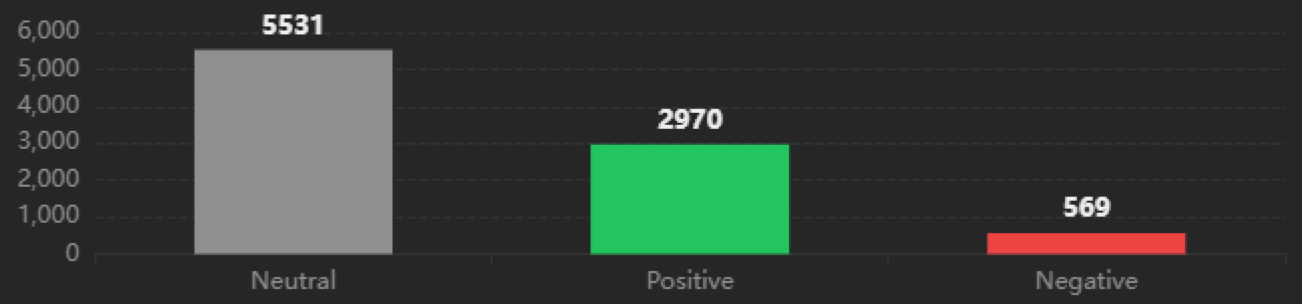
Origin	SXSW Conference Tweets
Total Tweets	9,093 (9,070 after cleaning)
Target Brands	Apple & Google Products
Labeling	Human-Annotated

Data Characteristics

- **Language:** English with social media slang, abbreviations, emojis
- **Noise:** URLs, user mentions (@), hashtags (#), typos
- **Complexity:** Sarcasm, context-dependent sentiment
- **Grammar:** Informal, inconsistent punctuation

Class Distribution

Challenge: Significant class imbalance requires specialized handling

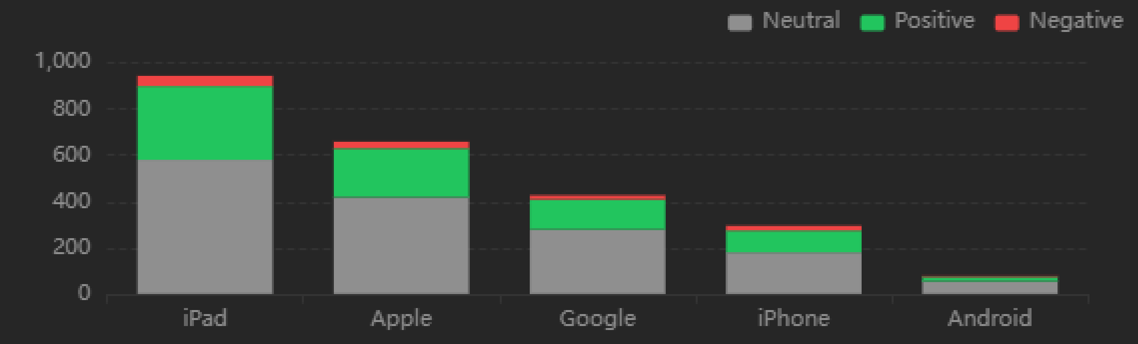


61%	33%	6%
Neutral	Positive	Negative
5,531 tweets	2,970 tweets	569 tweets

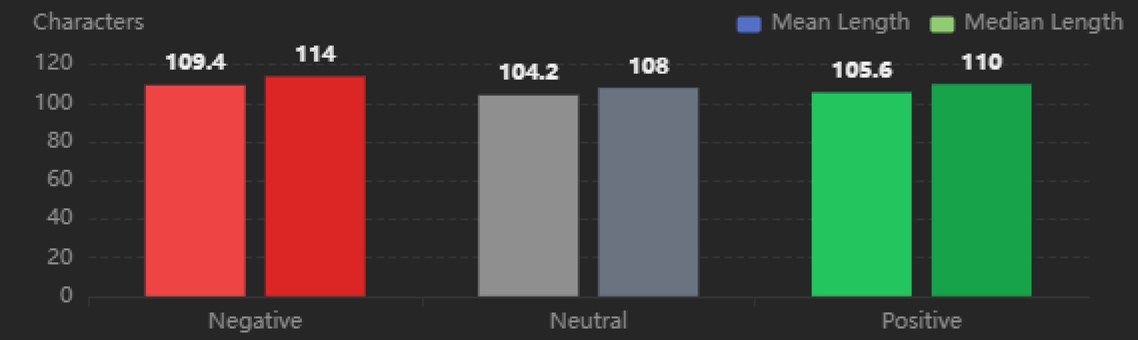
Implication: The 10:1 ratio between neutral and negative classes creates a significant modeling challenge. Standard accuracy metrics can be misleading—a model that always predicts "neutral" would achieve 61% accuracy while failing completely at crisis detection.

Exploratory Data Analysis Findings

Product Mention Analysis



Tweet Length by Sentiment



Key Insights

iPad Dominates Conversation

With 945 mentions, iPad is the most discussed product, followed by Apple (659) and Google (428). Most mentions are neutral, but there's a healthy proportion of positive sentiment.

Negative Tweets Are Longer

Negative tweets average 109 characters vs. 104 for neutral tweets. Complaints require more words to explain issues, while neutral tweets are often brief mentions.

SXSW Context Matters

The conference setting explains the 61% neutral class—high volume of promotional/informational tweets. Tech-savvy audience shows lower complaint rates typical of brand-loyal users.

Modeling Implication: Tweet length alone won't be a strong predictor—the specific words matter more than how many there are. Product-specific patterns suggest brand-level sentiment analysis could enhance performance.

Text Preprocessing Pipeline

Data cleaning is crucial for NLP. Poor quality data leads to poor model performance, regardless of model sophistication. Our comprehensive preprocessing pipeline transforms messy, unstructured text into clean, analyzable features.

Basic Cleaning

URL Removal
Strip http/https links from tweets

Mention Removal
Remove @username references

Hashtag Processing
Keep hashtag text, remove # symbol

Special Characters
Eliminate punctuation and numbers

Case Normalization
Convert all text to lowercase

Advanced NLP (NLTK)

Tokenization
Split text into individual words using NLTK word_tokenize

Lemmatization
Reduce words to root form using WordNet ("running" → "run")

Stopword Removal
Eliminate common words while preserving sentiment-bearing negations

Smart Retention
Keep critical sentiment words: "not", "no", "but", "against"

</> Before & After Examples

Original Tweet	After Basic Cleaning	Final Processed
"@apple I LOVE my new iPhone! Check out http://apple.com #awesome"	"i love my new iphone check out awesome"	"love new iphone check awesome"

Feature Engineering: TF-IDF & N-grams

TF-IDF Vectorization

Term Frequency-Inverse Document Frequency (TF-IDF) transforms text into numerical features by emphasizing distinctive words over common terms.

Max Features	N-gram Range
7,000 Most important terms	(1, 2) Unigrams + Bigrams

Min Doc Freq	TF Scaling
2 Eliminates rare words	Sublinear Logarithmic weighting

Why N-grams Matter

Individual words (unigrams) don't always capture meaning. Phrases like "not good" have the opposite meaning of "good" . Bigrams capture multi-word sentiment expressions that unigrams miss.

Top Bigrams by Sentiment

Positive Sentiment

apple store · sxsw link · come see · iphone app · popup store

Negative Sentiment

design headache · google circle · ipad design · crashy app

Neutral Sentiment

social network · new social · network called · google launch

Key Finding: Bigram analysis validates that capturing multi-word expressions significantly improves model performance by identifying sentiment-bearing phrases that unigrams miss entirely.

Model Development Strategy

We implemented an iterative approach, starting simple and progressively refining our models. Each iteration builds upon insights from the previous, systematically improving performance.

1 — Baseline
Naive Bayes

Fast baseline with interpretable probability estimates

Features	5,000
N-grams	(1,1)

2 — Enhanced
Enhanced NB

Added bigrams, increased vocabulary size

Features	7,000
N-grams	(1,2)

3 — Advanced
Logistic Reg

Handles feature dependence better than Naive Bayes

Weights	Balanced
Regularization	L2 (C=1.0)

4 — ⭑ SELECTED
Linear SVM

Optimal decision boundaries in high-dimensional space

Weights	Balanced
Regularization	C=0.5

5 — Ensemble
XGBoost

Gradient boosting with tree-based learning

Estimators	100
Learning Rate	0.1

Validation Strategy

Train-Test Split
80% training (7,255) / 20% testing (1,814)

Stratification
Preserves original sentiment distribution

Cross-Validation
5-fold CV for robustness assessment

Evaluation Metrics

Accuracy
Overall correctness

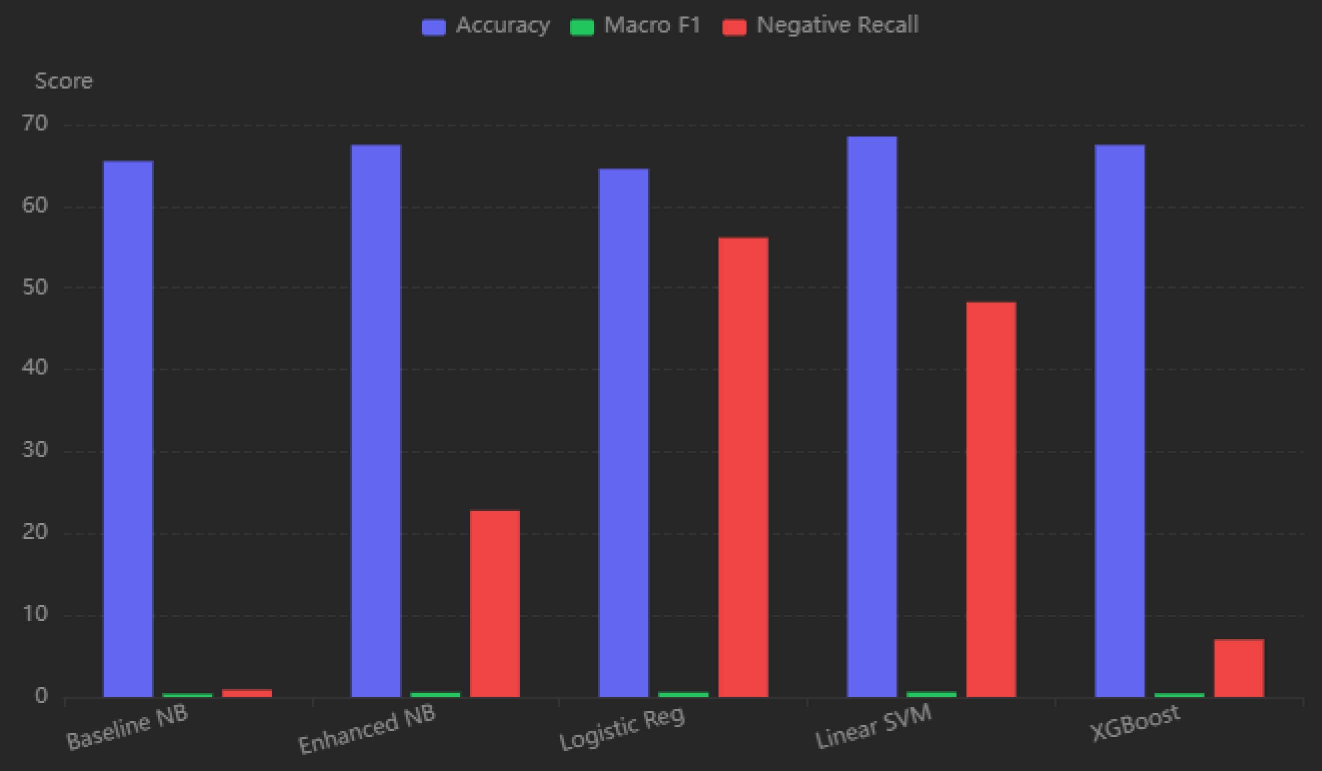
Precision
Predicted positives accuracy

Recall
Actual positives identified

F1-Score
Precision-recall balance

Model Performance Comparison

Comprehensive Model Comparison



Performance Matrix

Model	Acc	Macro F1	Neg Recall
Linear SVM ⋆	**68.52%**	**0.597**	**48.25%**
Enhanced NB	67.48%	0.541	22.81%
XGBoost	67.48%	0.442	**7.02%**
Baseline NB	65.49%	0.392	**0.88%**
Logistic Reg	64.55%	0.565	**56.14%**

Model Strengths

Linear SVM
Best overall balance, production-ready

Logistic Regression
Highest negative recall (56%)

Enhanced NB
Fast baseline, interpretable

The Accuracy Paradox

While XGBoost shows comparable accuracy (67.48%) to Linear SVM (68.52%), it misses **93% of negative sentiment** (7% recall). This demonstrates why accuracy alone is misleading for imbalanced datasets.

Linear SVM: Production Model Deep Dive

Linear SVM Selected

Best Overall Performance

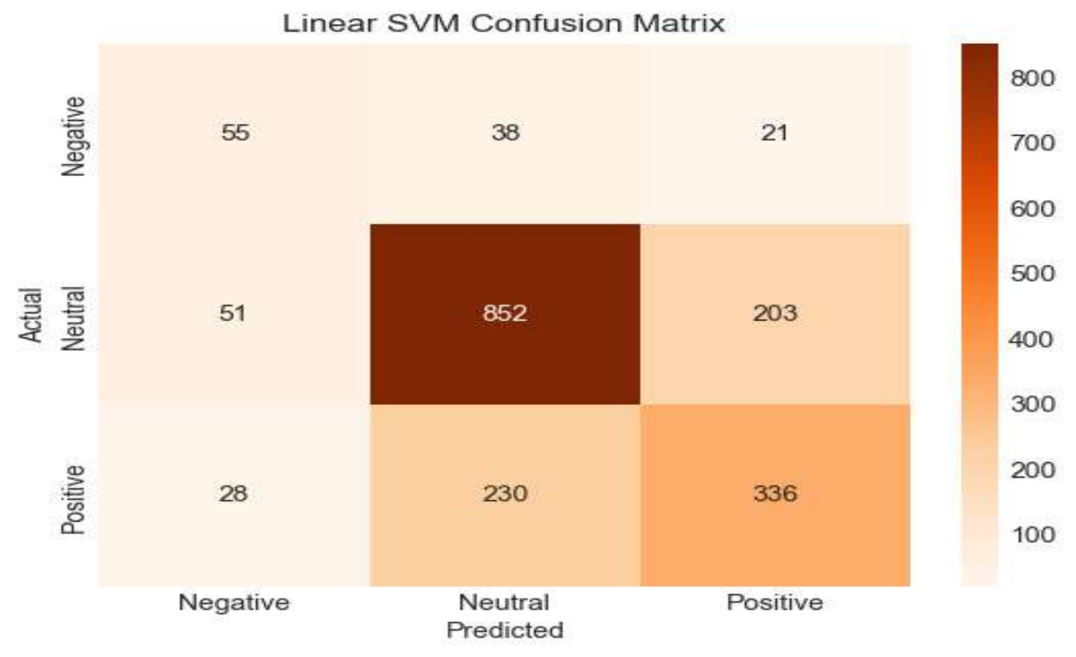
Overall Accuracy	Macro F1-Score
68.52%	0.597
Correct classification rate	Balanced performance

Negative Recall	Negative F1-Score
48.25%	0.444
Catches nearly half of complaints	Best balance for crisis detection

Why Linear SVM Won

- **Highest Overall Accuracy**
  68.52% ensures reliable predictions across all classes
- **Strong Negative F1-Score**
  0.44 balances recall and precision for crisis detection
- **Cross-Validation Stability**
  Mean F1: 0.554 ± 0.025 confirms stable generalization
- **Fewer False Positives**
  Reduces customer service workload vs. Logistic Regression

Confusion Matrix



- **Top Performance:** It led with 68.52% accuracy and the highest overall balance. It doesn't just guess; it understands language nuances.
- **Business Critical:** While others ignored complaints, the SVM delivered the highest Negative F1 (0.4435). It catches negative sentiment without constant false alarms.
- **Proven Stability:** 5-fold cross-validation confirmed a Mean F1 of 0.5542 with minimal variance. This model is reliable and ready for deployment.

Key Insights & Linguistic Patterns

Critical Findings

1 **The Accuracy Paradox**
XGBoost shows comparable accuracy (67.48%) to Linear SVM (68.52%) but misses **93% of negative sentiment** . Accuracy alone is misleading for imbalanced datasets.

2 **Sentiment Distribution Patterns**
iPad dominates (945 mentions, predominantly positive). Apple shows strong brand loyalty. iPhone has more polarized sentiment.

3 **Cross-Validation Stability**
Linear SVM shows low variance (±0.025) across 5-fold CV, confirming **stable generalization** to unseen data.

Business Implication: Linear SVM achieves the best business balance by maintaining high overall performance while detecting nearly half of all complaints.

Linguistic Patterns Discovered

Positive Sentiment Indicators

love | excited | amazing | check out | awesome

Action verbs and enthusiasm markers dominate positive tweets

Negative Sentiment Indicators

hope | issue | problem | not working | crash

Complaint markers and longer tweet length (109 vs 104 chars)

Neutral Sentiment Indicators

new social | google launch | network called | major new

Informational phrases without emotion words, shortest length

Model Trade-offs

Logistic Regression
Strength: Highest negative recall (56%)
Weakness: Lower overall accuracy

Linear SVM (Selected)
Strength: Best overall balance
Weakness: Moderate negative recall

Business Recommendations & Deployment

Immediate Deployment Actions

A. Operationalize Crisis Alerts

Set confidence threshold at 0.7 for high-priority alerts. Route negative tweets to specialized response dashboard with 15-minute SLA.

B. Customer Service Integration

Smart Routing: Auto-assign to product teams

Priority Queue: Surface high-confidence complaints

Context: Include product, length, confidence score

Sentiment Dashboard

Real-Time Pulse
Live sentiment gauge

Product Breakdown
By iPad, iPhone, etc.

Trend Analysis
Hourly/daily shifts

Top Issues
Common bigrams

Crisis Management Protocol

1. **Alert Level Yellow** — 20% Spike
Model detects spike in negative sentiment → Automated alert

2. **Alert Level Orange** — Confirmed
Human review confirms emerging issue → Escalate to PR team

3. **Alert Level Red** — Public
Public acknowledgment if trend continues → Full crisis mode

Product Launch Monitoring

Pre-Launch
Establish sentiment baseline 2 weeks before

Launch Day
Real-time monitoring with 5-minute refresh

Post-Launch
Track sentiment decay over 30 days

Expected Impact: Reduce crisis response time from hours to minutes, prioritize 50+ daily complaints automatically, and enable data-driven brand health monitoring.

Future Work & Project Conclusion

Model Improvement Roadmap

1 Phase 1: Short-Term (1-3 months)

- ✓ Feedback loop with customer service labels
- ✓ A/B test confidence thresholds (0.6 vs 0.7 vs 0.8)
- ✓ Error analysis on 100 misclassified tweets per class

2 Phase 2: Medium-Term (3-6 months)

- Emoji sentiment analysis (😊, 😢, 🔥)
- Aspect-based sentiment (battery, screen, price)
- SMOTE for class imbalance mitigation

3 Phase 3: Long-Term (6-12 months)

- BERT fine-tuning for 5-10% accuracy gain
- Multi-modal analysis (images + videos)
- Conversation context (tweet threads)

Project Success Summary

What We Achieved

- Built end-to-end NLP pipeline for sentiment classification
- Achieved 68.52% accuracy with robust negative detection
- Demonstrated the accuracy paradox in imbalanced datasets
- Identified key linguistic patterns for business insights

Key Takeaways

- Text preprocessing is critical for accurate predictions
- Model selection significantly impacts business outcomes
- Balanced metrics essential for imbalanced datasets
- Cross-validation ensures stable generalization

Business Applications

Monitor Opinion	User Behavior
Data Decisions	Brand Reputation

Thank You