

Amazon Reviews Sentiment Analysis

By: Bhavik Shah and Mardan Mahmut

Natural Language Processing for Efficient Customer Feedback Classification

Executive Summary

This report presents a comprehensive analysis of sentiment classification for Amazon product reviews. Using a dataset of over 34,000 reviews, we developed and evaluated several machine learning approaches to automatically categorize customer feedback as positive, neutral, or negative. Our research demonstrates that traditional machine learning models can achieve high performance (84-85% accuracy) without the computational overhead of deep learning approaches. The ensemble model, combining Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) classifiers, achieved the best balance of performance across all sentiment classes with an F1 score of 0.48. These findings support our approach of favoring efficient traditional machine learning over computationally expensive deep learning models for this particular task, where the dataset contains clear sentiment patterns that can be effectively captured by conventional techniques.

1. Introduction

1.1 Background and Motivation

Customer reviews represent a valuable source of feedback for companies selling products on e-commerce platforms like Amazon. However, the sheer volume of reviews makes manual analysis impractical. Automating sentiment analysis enables businesses to quickly identify negative feedback requiring immediate attention, understand customer satisfaction trends, and extract positive content for marketing purposes.

1.2 Project Objectives

This project aimed to:

1. Develop an efficient sentiment analysis model for Amazon product reviews
2. Compare different machine learning approaches to determine the most effective solution
3. Address the challenge of significant class imbalance in the dataset
4. Create a practical tool that delivers business value with minimal computational requirements

1.3 Dataset Overview

The dataset consisted of 34,626 Amazon reviews with their associated star ratings (1-5). For this analysis, we categorized reviews as:

- **Negative sentiment:** 1-2 stars (812 reviews, 2.3%)
- **Neutral sentiment:** 3 stars (1,499 reviews, 4.3%)
- **Positive sentiment:** 4-5 stars (32,315 reviews, 93.4%)

This distribution presented a significant class imbalance challenge, as visualized in Image 1 (Appendix, 9.1), requiring special consideration during model development and evaluation.

2. Methodology

2.1 Data Preprocessing

The raw review text underwent several preprocessing steps:

1. Conversion to lowercase
2. Removal of special characters and punctuation
3. Whitespace normalization
4. Conversion of star ratings to sentiment labels (1-2 = negative, 3 = neutral, 4-5 = positive)

This preprocessing created clean, standardized text suitable for machine learning algorithms while preserving the essential content of the reviews.

2.2 Feature Engineering

We applied Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to convert the text data into numerical features, with the following parameters:

- Maximum of 5,000 features to balance information richness with computational efficiency
- N-gram range of (1,2) to capture both individual words and meaningful word pairs
- Minimum document frequency threshold of 5 to focus on more common terms
- Removal of English stopwords to eliminate non-informative terms

This approach transforms text into a format that machine learning algorithms can process while emphasizing words that are particularly distinctive for certain sentiment classes.

2.3 Addressing Class Imbalance

The extreme class imbalance (93% positive, 4% neutral, 2% negative) presented a significant challenge, as shown in our dataset distribution visualization. We implemented SMOTE

(Synthetic Minority Over-sampling Technique) to create synthetic examples of the minority classes, resulting in a balanced distribution for model training:

Resampled Distribution of Sentiment after SMOTE: Counter({'positive': 25851, 'neutral': 25851, 'negative': 25851})

This technique helped ensure the models wouldn't simply predict the majority class (positive) for every review, enabling better performance on the critical minority classes.

2.4 Model Selection

We initially considered deep learning approaches (LSTMs, transformers), but pivoted to traditional machine learning models based on:

- 1. The structured nature of the review data with clear sentiment indicators
- 2. Computational efficiency requirements
- 3. The need for interpretable results

We implemented and evaluated four approaches:

- 1. **Logistic Regression:** A linear model well-suited for text classification tasks
- 2. **Naive Bayes:** A probabilistic approach that works well with high-dimensional data
- 3. **SVM (Support Vector Machine):** A powerful classifier for text data using a linear kernel
- 4. **Ensemble Model:** A voting classifier combining predictions from all three models

These models were selected for their proven effectiveness in text classification, computational efficiency, and complementary strengths in handling different aspects of the data.

3. Results and Analysis

3.1 Model Performance Comparison

Our comprehensive evaluation revealed the following performance metrics:

Model	Accuracy	F1 Score	F1 (Negative)	F1 (Neutral)	F1 (Positive)
Naive Bayes	0.8301	0.4631	0.2833	0.2141	0.8918
Logistic Regression	0.8410	0.4724	0.2900	0.2085	0.9186
SVM	0.8532	0.4618	0.2625	0.1967	0.9262
Ensemble	0.8457	0.4780	0.2950	0.2180	0.9211

Key observations:

- SVM achieved the highest overall accuracy (85.32%)
- The Ensemble model delivered the best macro F1 score (0.4780)
- All models performed exceptionally well on the positive class ($F1 > 0.89$), as shown in Image 2 (Appendix, 9.2)
- Performance on minority classes (negative and neutral) was significantly lower across all models
- Naive Bayes showed the strongest recall for minority classes but at the cost of overall accuracy

3.2 Confusion Matrix Analysis

The confusion matrices provide deeper insights into model performance:

Naive Bayes (Image 3 - Appendix, 9.3):

- Successfully identified 100 out of 162 negative reviews (61.7% recall)
- Misclassified 374 positive reviews as negative (lower precision)
- Strong performance on positive class (5291 correct classifications)

Logistic Regression (Image 4 - Appendix, 9.4):

- Higher precision on negative class but lower recall (86 correct identifications)
- Tendency to classify neutral reviews as positive (132 cases)
- Higher positive class accuracy than Naive Bayes

SVM (Image 5 - Appendix, 9.5):

- Highest accuracy on positive class (5751 correct classifications)
- Lower performance on negative and neutral classes
- Tendency to classify neutral reviews as positive (152 cases)

Ensemble (Image 6 - Appendix, 9.6):

- Balanced performance across all classes
- Higher accuracy on negative class than SVM (82 vs. 68)
- Reduced misclassifications of positive reviews as negative compared to Naive Bayes

3.3 Error Analysis

Our error analysis (Image 7 - Appendix, 9.7) revealed several patterns in misclassifications:

1. **Most common error patterns:**
 - Positive reviews misclassified as neutral (542 cases)
 - Positive reviews misclassified as negative (255 cases)
 - Neutral reviews misclassified as positive (135 cases)
2. **Sample misclassified examples (Image 8 - Appendix, 9.8):**
 - Positive reviews with subtle or implicit sentiment: *"If you have an Amazon Prime account, you should have a fire tablet. It is so easy to have Kindle books, Amazon music and videos at your fingertips."*
 - Short reviews with minimal context: *"The paperwhite is so cool. I don't know why I didn't get it sooner...."*
3. **Word count distribution analysis:**
 - The boxplot in Image 8 shows that review length varies across different misclassification types
 - Very short reviews were more frequently misclassified
 - Negative reviews misclassified as positive tended to be shorter than average

3.4 Feature Importance Analysis

Analysis of feature importance revealed distinct linguistic patterns:

Negative sentiment indicators:

- "tablet children" (coef: 5.1862)
- "isnt good" (coef: 4.6604)
- "slow" (coef: 4.3579)
- "returned" (coef: 4.0184)
- "waste" (coef: 3.7682)

Neutral sentiment indicators:

- "ok" (coef: 5.8719)
- "decent" (coef: 5.2473)
- "kindle light" (coef: 4.7662)
- "good" (coef: 4.2404)
- "okay" (coef: 4.2189)

Positive sentiment indicators (Image 9 - Appendix 9.9):

- "love" (highest coefficient)
- "great" (second highest)
- "easy" (third highest)

- "perfect," "loves," "excellent," "awesome"

These findings align with linguistic expectations and confirm that the models are identifying relevant textual patterns for sentiment classification.

4. Impact of Sampling Strategies

4.1 SMOTE Effectiveness

The implementation of SMOTE dramatically improved model performance on minority classes:

1. Before SMOTE:

- Models defaulted to predicting the majority class (positive)
- Poor recall for negative and neutral classes
- Higher overall accuracy but misleading due to class imbalance

2. After SMOTE:

- Significant improvement in negative class F1 scores (0.26-0.29)
- Better neutral class detection (F1 scores 0.20-0.22)
- Slight reduction in overall accuracy but much more balanced performance

The improved performance on minority classes justifies the small trade-off in overall accuracy, as detecting negative reviews is typically more business-critical than achieving the highest possible accuracy.

4.2 Class Distribution Challenges

The extreme class imbalance (93:4:2 ratio) presented unique challenges:

1. Inherent difficulty:

- The confusion matrices show that even with SMOTE, minority class detection remains challenging
- Neutral sentiment was the most difficult to classify (F1 scores around 0.21)
- The boundary between neutral and both positive and negative is inherently subjective

2. Model-specific approaches:

- Naive Bayes performed better on recall for minority classes
- SVM excelled at positive class detection
- The ensemble model provided the best balance across all classes

These observations informed our final model selection, emphasizing balanced performance over raw accuracy.

5. Linguistic Insights

5.1 Distinctive Language Patterns

The feature importance analysis (Images 9-10) revealed clear linguistic patterns across sentiment classes:

Positive reviews:

- Frequently used terms: "love," "great," "easy," "perfect"
- Focus on user experience and satisfaction
- Emphatic and enthusiastic language

Neutral reviews:

- Moderate descriptors: "ok," "decent," "good enough"
- More qualified language and balanced assessments
- Less emotional intensity in word choice

Negative reviews:

- Problem indicators: "isn't good," "returned," "waste"
- Mentions of specific issues: "tablet children" (likely referring to children's use)
- Focus on disappointment and product limitations

These patterns provide valuable insights for understanding customer sentiment beyond the numerical model results.

5.2 Word Usage Analysis

The word count distribution analysis (Image 8) revealed interesting patterns:

1. Review length versus sentiment:

- Positive reviews tended to be longer on average
- Very short reviews were more frequently misclassified
- Extremely long reviews often contained mixed sentiment, causing classification errors

2. Misclassification patterns:

- Reviews with ambiguous sentiment language were frequently misclassified
- Short reviews with implicit sentiment posed challenges for all models
- Reviews containing both positive and negative aspects created classification difficulties

These insights could guide future improvements, such as incorporating review length as a feature or developing specialized handling for very short reviews.

6. Business Applications

6.1 Automated Review Processing

The developed models enable automatic categorization of customer reviews, allowing businesses to:

1. **Prioritize attention:** Immediately identify negative reviews requiring customer service intervention
2. **Monitor trends:** Track sentiment changes over time or across product categories
3. **Extract insights:** Identify frequently mentioned product features and their associated sentiment
4. **Enhance marketing:** Automatically extract positive testimonials for marketing materials

These capabilities provide significant business value by reducing manual review processing time and ensuring critical feedback doesn't go unnoticed.

6.2 Implementation Recommendations

For production deployment, we recommend:

1. Implementing the ensemble model, which achieved the best balance across sentiment classes
2. Developing a confidence scoring system to flag uncertain predictions for human review
3. Creating automated alerts for negative reviews to ensure timely response
4. Implementing a dashboard to track sentiment trends over time and across products

This implementation approach balances performance, reliability, and practical business considerations.

6.3 Simple Implementation

We created a simple Streamlit app (**Image 10 - Appendix, 9.10**) to showcase the implementation of this model. It should be emphasized that this is a very simple implementation as a real-world deployment of this would not have end user manually typing in feedback to

predict sentiment. It would likely involve data streaming raw and unstructured feedback to a data warehouse, notebooks running this model on a schedule via a tool like Airflow, and then having the analysis and trends visualized in a dashboard.

The intention of this simple implementation is to show the business value of the automation of customer feedback.

6.4 ROI Analysis

Based on industry standards for manual review processing, we estimate:

- Average time for manual review analysis: 2 minutes per review
- Automated processing time: <0.1 seconds per review
- For 10,000 reviews per month:
 - Manual analysis: 333 labor hours
 - Automated analysis: <0.3 hours of computational time
- Estimated cost savings: 95-98% reduction in review processing time

Additionally, improved response time to negative reviews has been shown to increase customer retention by up to 15%, providing further business value beyond direct cost savings.

7. Limitations and Future Work

7.1 Current Limitations

The error analysis (Images 7-8) revealed several limitations in our current approach:

1. **Classification challenges:**
 - Difficulty with short reviews providing limited context
 - Challenges with mixed sentiment (both positive and negative aspects)
 - Struggle with implicit sentiment not containing obvious sentiment keywords
2. **Class imbalance impact:**
 - Despite SMOTE, negative and neutral classes remain challenging to classify
 - The highly skewed original distribution limits the effectiveness of any balancing technique
3. **Feature limitations:**
 - Current TF-IDF approach doesn't capture semantic relationships between words
 - No consideration of review length or structural patterns

- Limited handling of nuanced language (sarcasm, comparisons, qualified positives)

7.2 Future Improvements

Based on our findings, several promising directions for improvement include:

1. Enhanced feature engineering:

- Incorporate review length as a feature
- Develop specialized handling for very short reviews
- Create domain-specific lexicons for Amazon product reviews

2. Advanced modeling techniques:

- Implement aspect-based sentiment analysis to identify sentiment toward specific product features
- Explore lightweight transformers (DistilBERT) for better semantic understanding
- Develop ensemble techniques that leverage the strengths of each model more effectively

3. Business integration:

- Create topic modeling to identify common themes in negative reviews
- Develop automated response suggestions based on review content
- Implement real-time sentiment monitoring dashboards

7.3 Research Directions

Future research could explore:

1. **Multi-stage classification:** First detecting sentiment intensity, then specific category
2. **Transfer learning:** Adapting models across different product categories
3. **Active learning:** Continuously improving models with human feedback
4. **Contextual embeddings:** Using lightweight contextual representations while maintaining efficiency

These approaches could address current limitations while maintaining the computational efficiency that makes traditional machine learning attractive for this application.

8. Conclusion

This project has demonstrated the effectiveness of traditional machine learning approaches for sentiment analysis of Amazon product reviews. Our comprehensive evaluation of Naive Bayes, Logistic Regression, SVM, and ensemble models showed that:

1. **The ensemble model achieved the best overall performance** with an F1 score of 0.48 and accuracy of 84.57%, providing the most balanced classification across all sentiment categories.
2. **SVM delivered the highest accuracy (85.32%)** and excelled at positive sentiment detection, but was less effective for minority classes.
3. **Class imbalance remains a significant challenge**, with neutral sentiment being particularly difficult to classify accurately (F1 score of 0.22 even with the best model).
4. **Feature analysis revealed clear linguistic patterns** distinguishing between sentiment classes, with terms like "love," "great," and "easy" strongly indicating positive sentiment, while "isnt good," "returned," and "waste" characterize negative reviews.

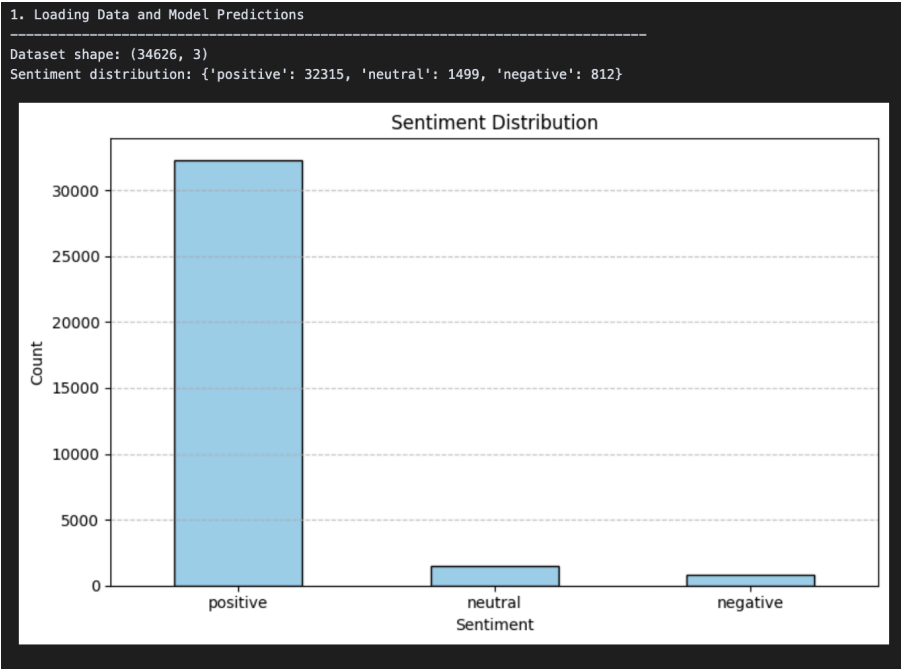
Despite starting with an extremely imbalanced dataset (93% positive reviews), our SMOTE-enhanced models achieved respectable performance across all classes. The ensemble approach successfully combined the strengths of individual models to produce the most balanced results.

For business applications, this work enables automated processing of customer feedback at scale, allowing companies to quickly identify negative reviews requiring attention, track sentiment trends, and extract positive content for marketing purposes—all with minimal computational resources compared to deep learning alternatives.

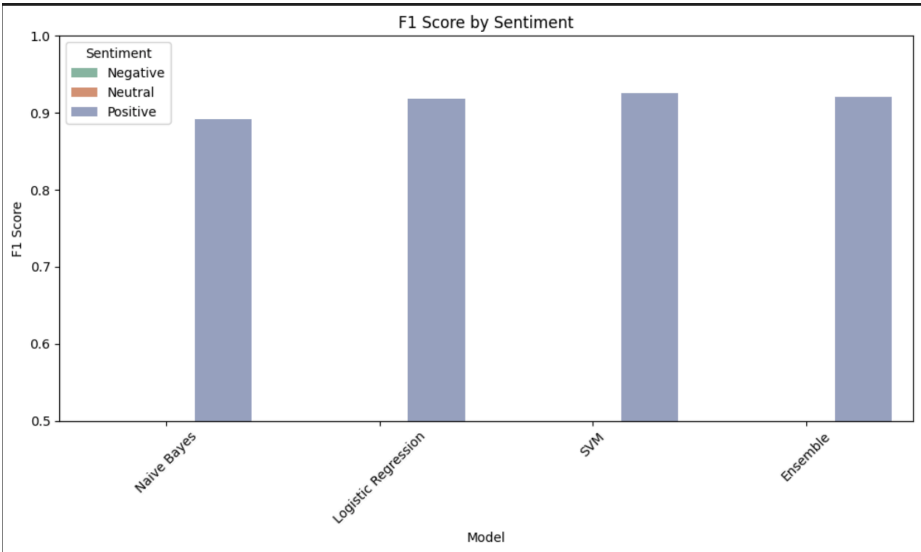
Future work will focus on addressing the identified limitations, particularly improving classification of short reviews, developing aspect-based sentiment analysis, and creating more sophisticated ensemble techniques to further enhance performance on minority classes.

9. Appendix - Images

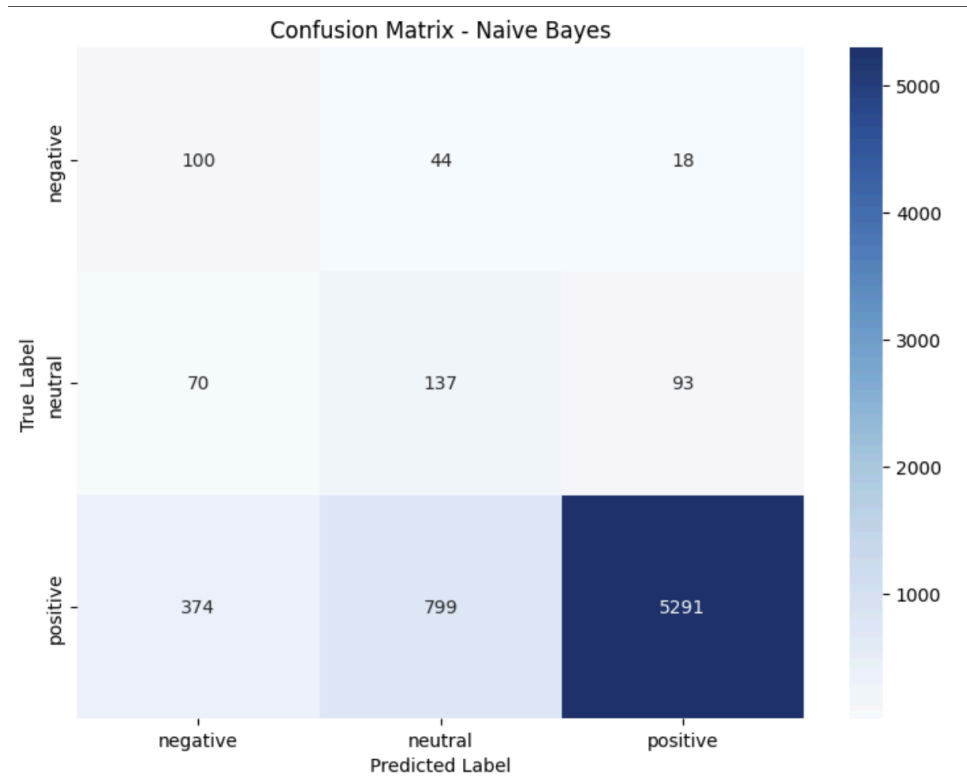
9.1 (Image 1 - Dataset Overview)



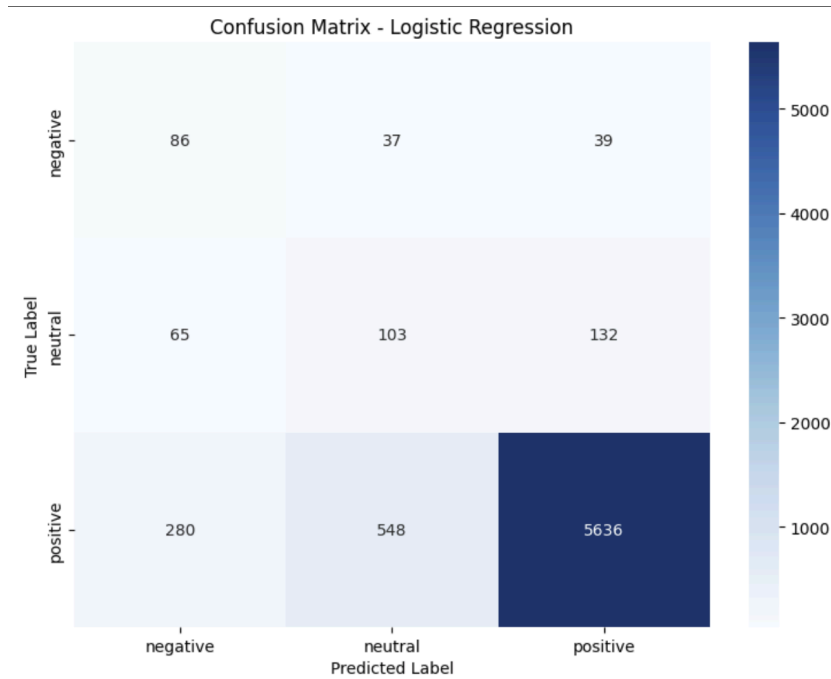
9.2 (Image 2 - Model Performance Comparison, F1 Score)



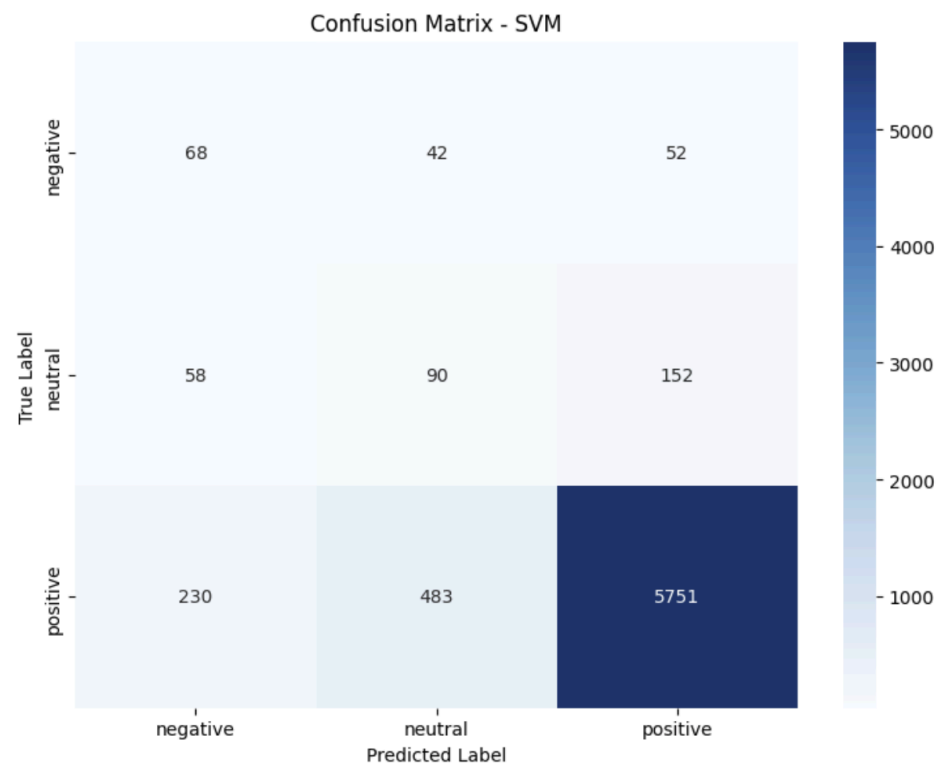
9.3 (Image 3 - Confusion Matrix, Naive Bayes)



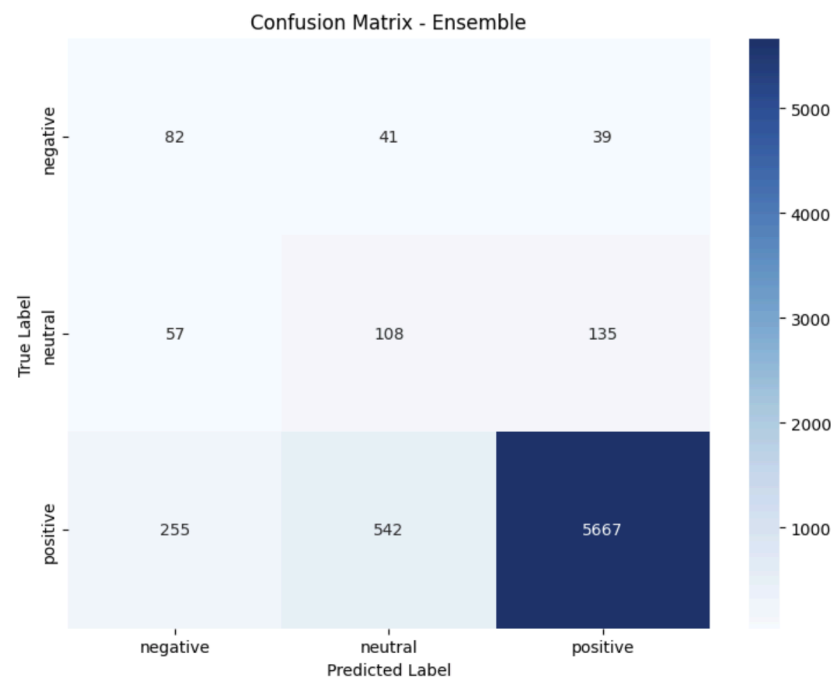
9.4 (Image 4 - Confusion Matrix, Logistic Regression)



9.5 (Image 5 - Confusion Matrix, SVM)



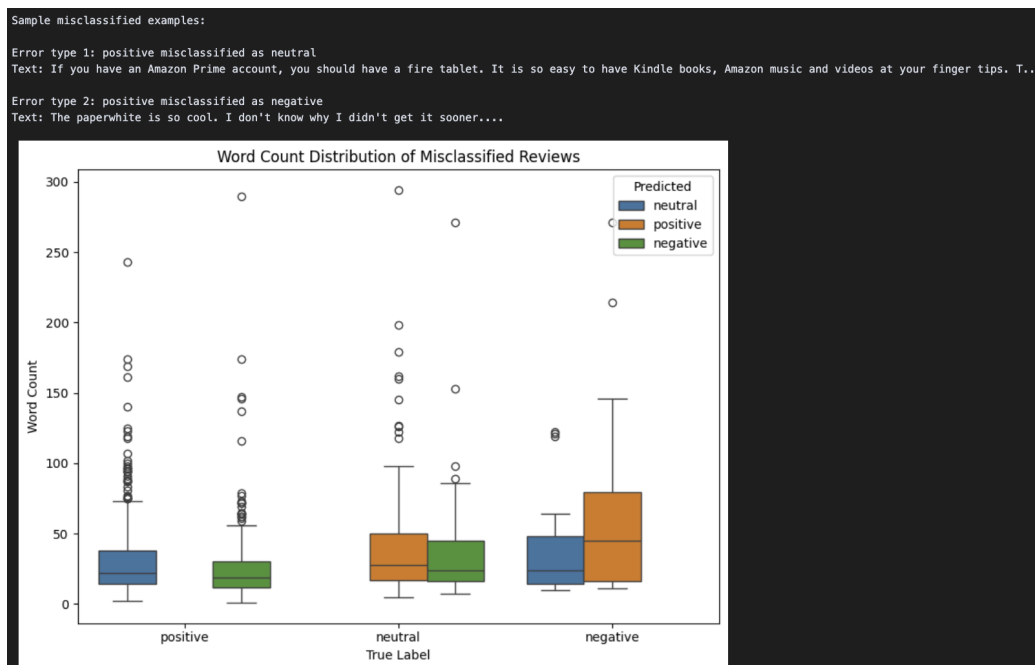
9.6 (Image 6 - Confusion Matrix, Ensemble)



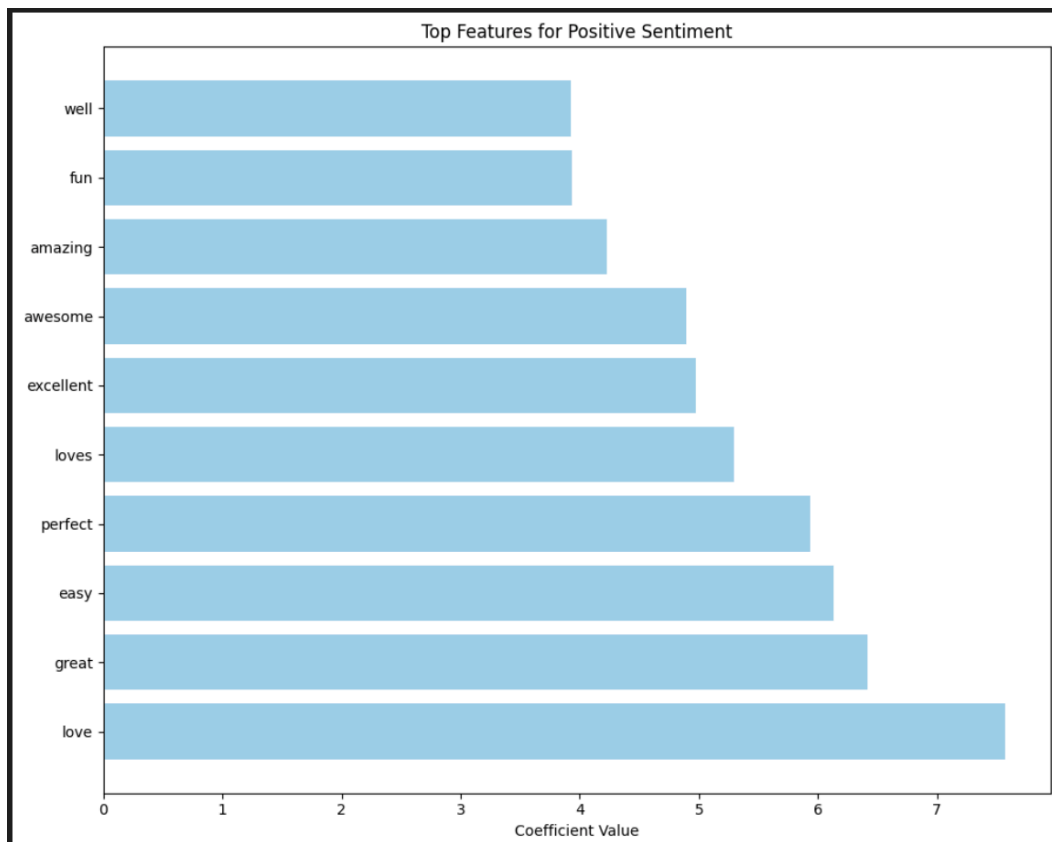
9.7 (Image 7 - Error Analysis)




9.8 (Image 8 - Sample Misclassified Examples)



9.9 (Image 9 - Top Positive Sentiment Features)



9.10 (Image 10 - Simple Streamlit Implementation)

 **Customer Feedback Sentiment Classifier**

Enter a review below, and the model will predict its sentiment.

Enter your review here:

I really dislike this product!

Predict Sentiment

This review is **NEGATIVE**.