Customer Review Sentiment Classifier

By Abhinaya Gyawali

Problem & Context

Understanding Customer Sentiment at Scale

1. Businesses receive large volumes of unstructured customer reviews, making manual analysis slow, inconsistent, and impractical.

2. Understanding customer sentiment is critical for product improvement, customer experience, and marketing decisions, but human review does not scale.

3. This project builds an NLP-based sentiment classifier that categorizes Amazon customer reviews into positive, neutral, or negative sentiment.

4. Automated sentiment classification enables faster, data-driven decision-making by transforming raw text feedback into structured insights.

Defining Success, Stakeholders, and Constraints

Evaluation criteria, business users, and practical limitations shaping the solution

Success Metrics	• ≥85% overall accuracy • Macro F1-score ≥ 0.80 (balanced class performance) • Scalable processing of large review volumes
Stakeholders	• Product Managers • Customer Success & Support Teams • Marketing & Business Analysts • Executive Leadership
Constraints	• Noisy and imbalanced text data • Limited time and compute resources • Difficulty capturing sarcasm and ambiguous sentiment • Limited generalization beyond Amazon reviews

Formulating the Problem

Translating business needs into a supervised NLP classification problem

- The business problem is framed as a supervised multi-class text classification task, where customer review text is used to predict sentiment labels: positive, neutral, or negative.
- Input features consist of unstructured natural language text extracted from customer reviews, while target labels are derived from star ratings mapped to sentiment classes.
- The objective is to train and evaluate NLP models that can accurately and consistently classify sentiment, enabling scalable analysis of customer feedback.
- Model performance is assessed using accuracy and macro F1-score to ensure balanced prediction quality across all sentiment classes.

Dataset Characteristics

Understanding Data Sources and Quality

- The project uses the Amazon US Customer Reviews dataset, which contains large-scale, real-world customer feedback across multiple product categories.
- Each record includes review text, star ratings, and supporting metadata such as product and review identifiers.
- Star ratings are converted into sentiment labels (positive, neutral, negative) to create a supervised learning dataset.
- The dataset reflects natural language variability, including differences in writing style, length, and sentiment intensity, making it suitable for evaluating real-world NLP performance.

Data Wrangling

Preparing raw customer reviews for reliable sentiment modeling

Filtered and selected relevant fields, focusing on review text and star ratings for sentiment analysis.

Converted star ratings into sentiment labels (positive, neutral, negative) to create a supervised learning target.

Cleaned review text by lowercasing, removing noise (special characters, formatting issues), and handling missing or invalid entries.

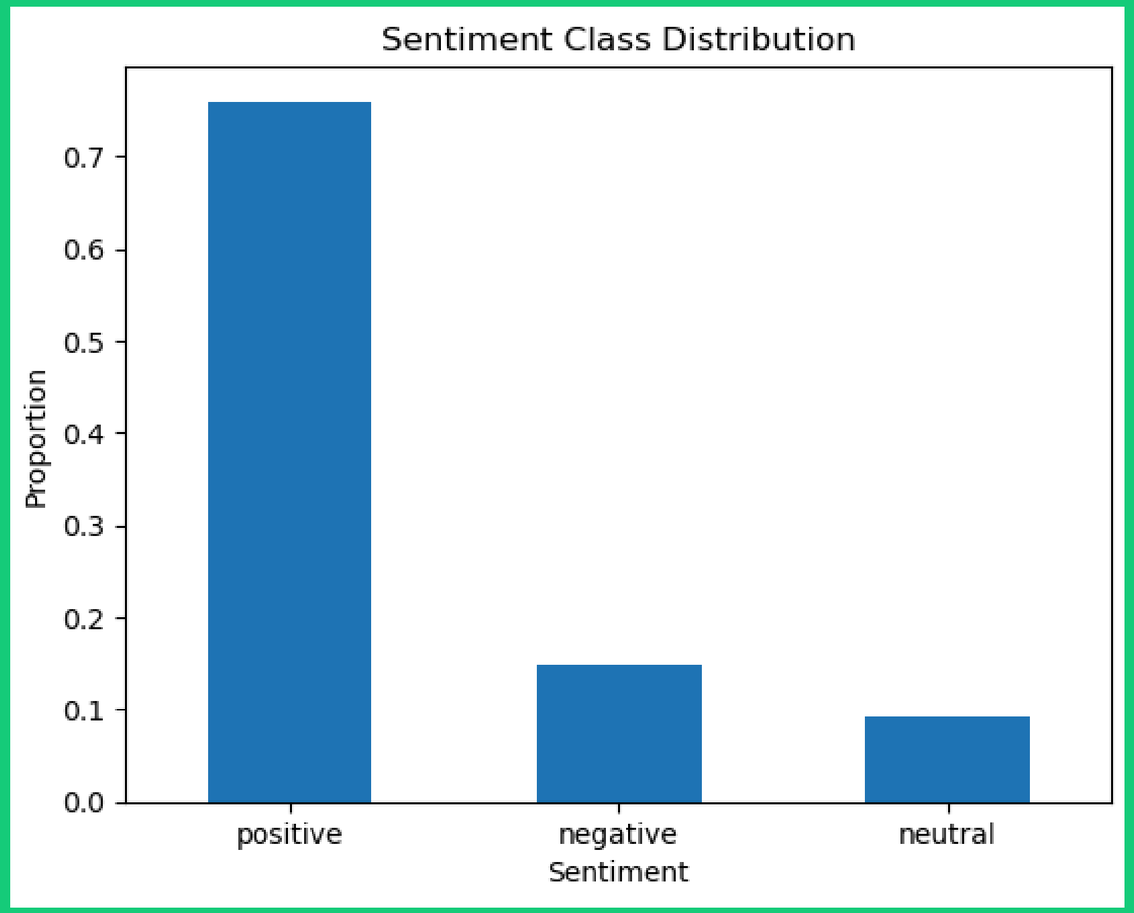
Checked for class imbalance, duplicates, and text length variability to understand data quality before modeling.

Prepared the dataset for downstream NLP pipelines, ensuring consistency across training and evaluation splits.

Key Data Insights from Wrangling

What the cleaned data reveals about sentiment classification challenges

- Customer reviews show significant variation in length and writing style, reinforcing the need for robust text preprocessing.
- Sentiment classes are imbalanced, with positive reviews dominating—highlighting the importance of macro F1-score over accuracy alone.
- Neutral sentiment is less clearly defined, often overlapping linguistically with positive and negative reviews.
- Real-world noise (informal language, ambiguity) confirms that sentiment classification is a non-trivial NLP task, even with labeled data.

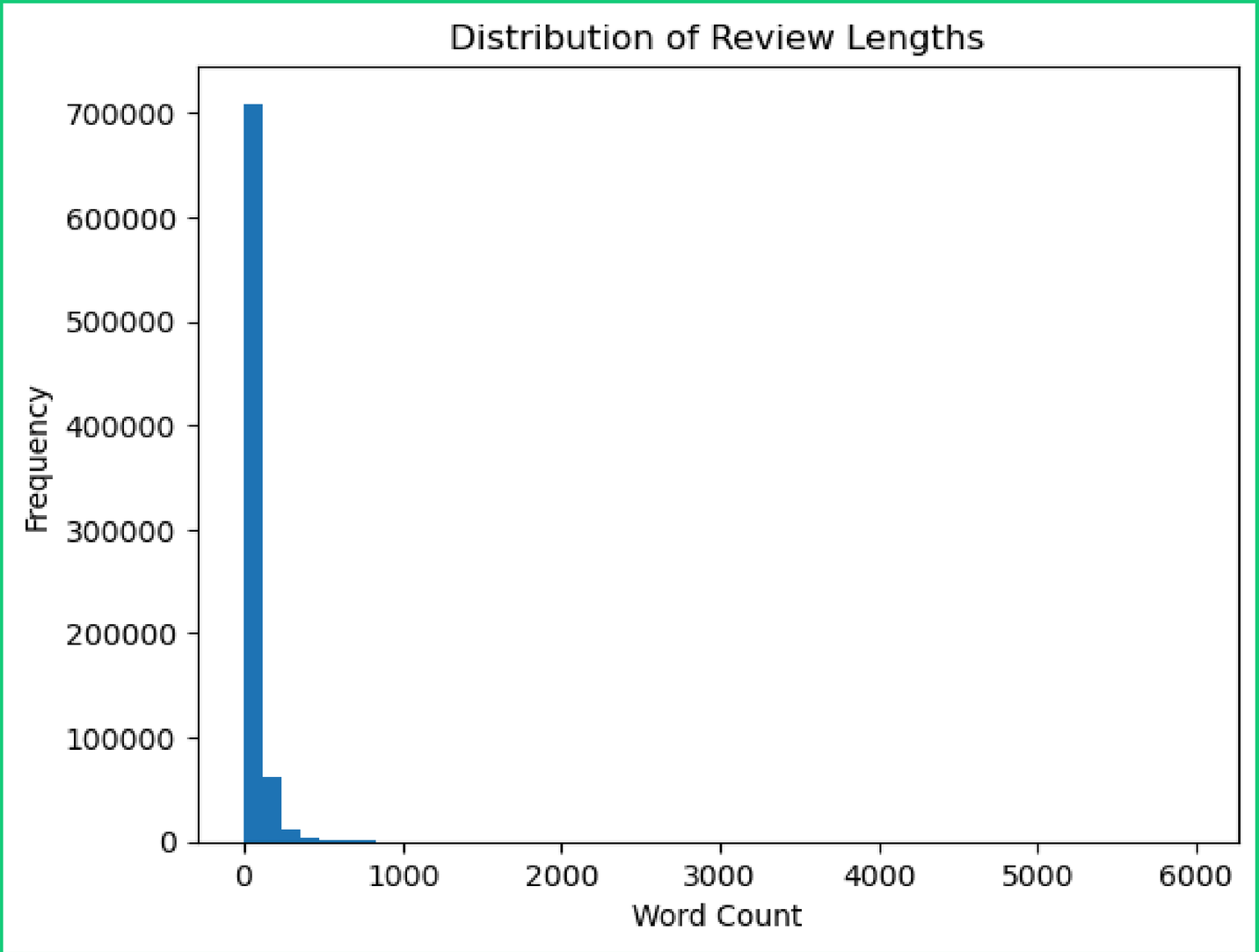


Sentiment Class Distribution

Positive reviews dominate the dataset, making class imbalance a key challenge and requiring balanced evaluation metrics such as macro F1-score.

Distribution of Review Lengths

Review lengths are highly skewed, with most reviews being short and a small number of very long reviews, reinforcing the need for robust text preprocessing and feature extraction.



Modeling Approach

Evaluating multiple NLP classification models for sentiment prediction

- The problem was formulated as a multi-class supervised text classification task, predicting positive, neutral, and negative sentiment from customer review text.
- Review text was converted into numerical features using TF–IDF vectorization, a proven method for representing textual importance in document classification.
- Multiple models were trained and compared, including Naive Bayes, Logistic Regression, and Linear SVM.
- This comparative approach ensured that model selection was driven by data performance rather than model complexity.

Model Performance Comparison

TF–IDF + Logistic Regression achieved the best balance between accuracy and macro F1-score, outperforming Naive Bayes and matching Linear SVM while offering greater interpretability.

Model	Test Samples	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted Precision	Weighted Recall	Weighted F1
Logistic Regression (TF-IDF)	158175	0.88	0.75	0.67	0.69	0.86	0.88	0.87
Linear SVM (TF-IDF)	158175	0.88	0.76	0.66	0.67	0.86	0.88	0.86
Naive Bayes (TF-IDF)	158175	0.85	0.72	0.6	0.62	0.83	0.85	0.83

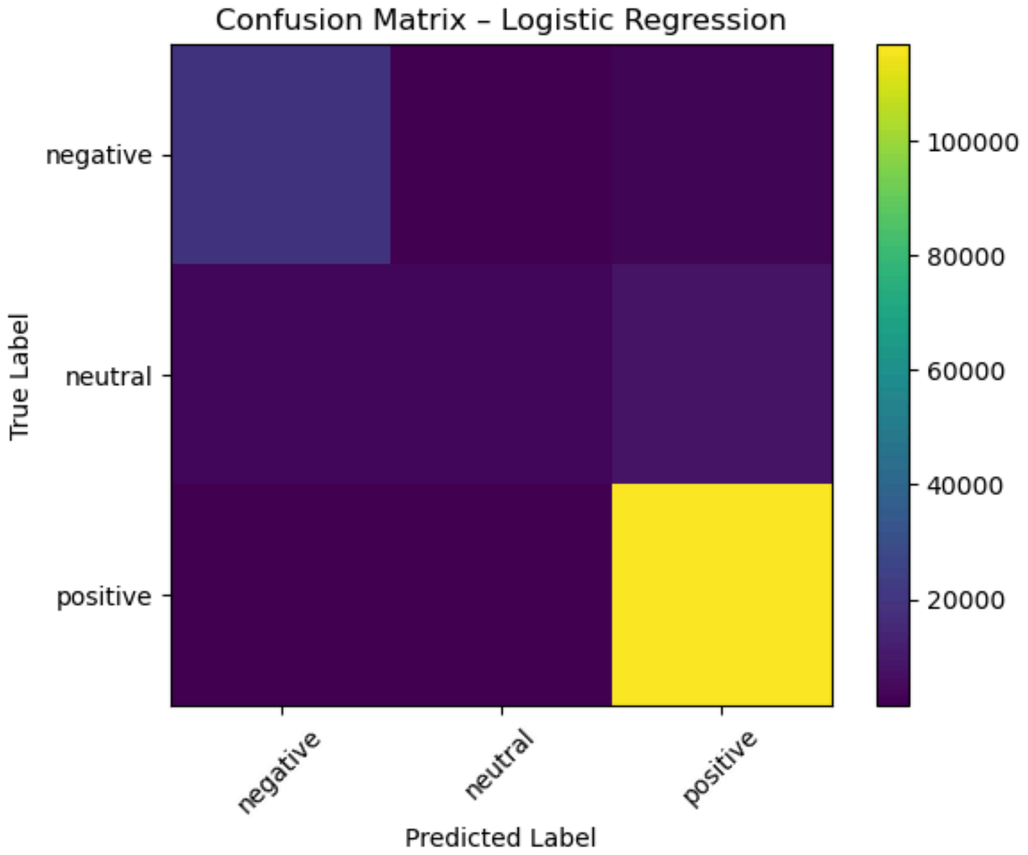
Final Model Selection & Rationale

Why TF–IDF + Logistic Regression was selected

- Logistic Regression with TF–IDF features was selected as the final model based on its strong macro F1-score and stable performance across all sentiment classes.
- The model effectively handled class imbalance, particularly improving performance on neutral and negative reviews compared to simpler baselines.
- Logistic Regression provides interpretability, allowing analysis of influential terms driving sentiment predictions.
- Compared to transformer-based models, it achieved competitive performance with significantly lower computational cost, making it practical for scalable deployment.

Confusion Matrix: Logistic Regression

The model predicts positive sentiment most accurately, while neutral sentiment shows higher misclassification due to overlap with positive and negative language.



Client Recommendations

How sentiment insights can be used to drive business impact

- Prioritize negative sentiment monitoring to quickly identify recurring product or service issues and reduce customer dissatisfaction.
- Track sentiment trends over time to measure the impact of product updates, marketing campaigns, or operational changes.
- Leverage positive reviews to identify key strengths and inform marketing messaging and customer testimonials.
- Flag neutral reviews for deeper analysis, as they often contain mixed feedback that can highlight subtle improvement opportunities.

Practical Considerations & Future Work

Limitations, enhancements, and next steps for real-world deployment

- Neutral sentiment remains challenging due to linguistic overlap with positive and negative reviews, impacting classification precision.
- Model performance may be affected by sarcasm, slang, and informal language, which are difficult to capture with traditional text features.
- Future improvements include fine-tuning transformer-based models to better capture semantic context and nuanced sentiment.
- Expanding training data to include reviews from additional platforms would improve generalizability.
- Incorporating human-in-the-loop feedback could continuously refine model predictions over time.

Thank you!