# E-commerce Sentiment Intelligence

A comprehensive sentiment analysis and customer intelligence system for Amazon Electronics reviews, implementing multi-model machine learning approaches for sentiment classification, customer segmentation, and predictive analytics.

**Executive Summary:-**

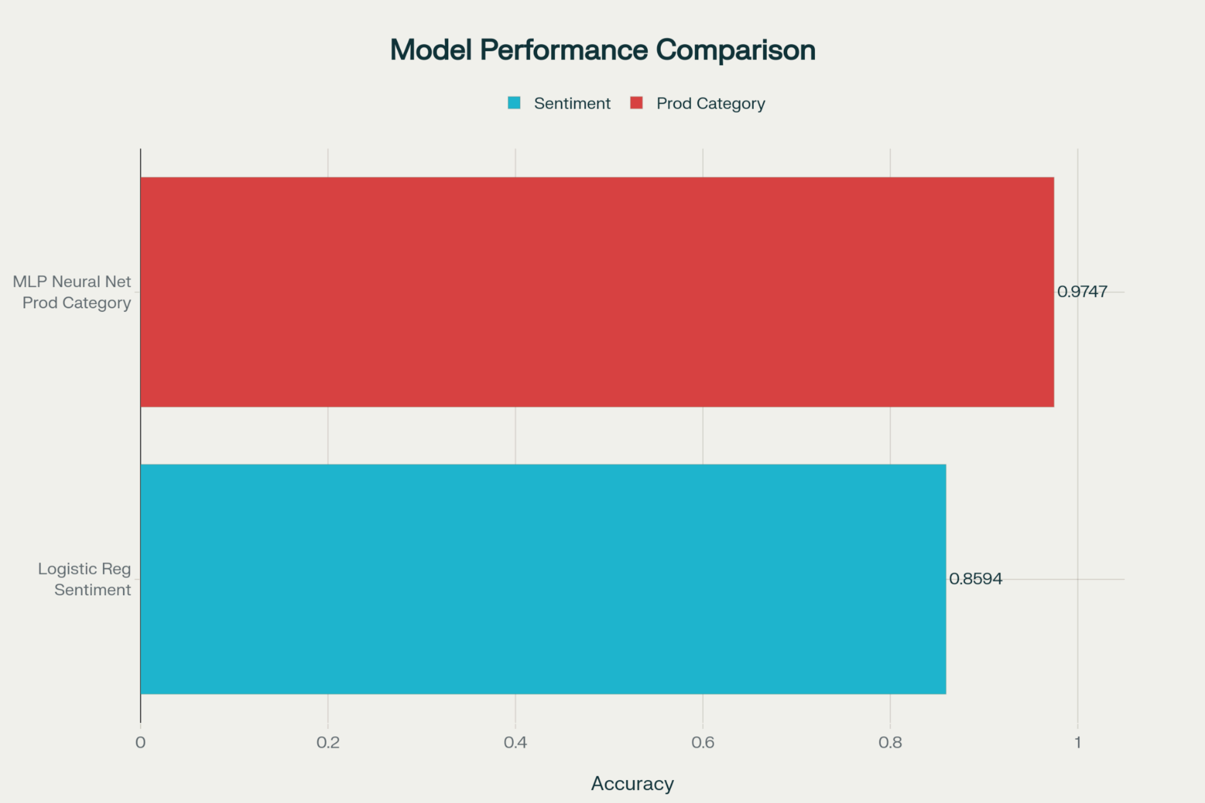
This project develops an end-to-end sentiment intelligence system for e-commerce platforms, analyzing over 1.5 million Amazon Electronics reviews from 2010-2014. Building upon foundational work in sentiment analysis, this implementation demonstrates the practical application of natural language processing and machine learning techniques to extract actionable business insights from customer feedback.

The system achieved 85.94% accuracy in sentiment classification using Logistic Regression, successfully identified five distinct customer segments through unsupervised clustering, and demonstrated 97.47% accuracy in predicting product categories from review text. These results validate the effectiveness of combining multiple analytical approaches- supervised classification, unsupervised clustering, and neural network-based prediction, to create a comprehensive understanding of customer sentiment and behavior patterns.

**Key Accomplishments:**

- Processed and analyzed 1,494,070 customer reviews with sophisticated text preprocessing pipelines
- Implemented three-class sentiment classification (Positive, Neutral, Negative) with strong performance on majority classes
- Discovered meaningful customer segments corresponding to distinct product categories (tablets, photography, storage, connectivity)
- Extracted interpretable features revealing sentiment drivers such as "great," "excellent," and "perfect" for positive sentiment versus "useless," "waste," and "worst" for negative sentiment

This work demonstrates the value of sentiment intelligence systems for e-commerce platforms seeking to understand customer opinions, identify market segments, and predict consumer preferences at scale.



Model performance comparison showing accuracy of Logistic Regression for sentiment classification and MLP Neural Network for product category prediction.

### Methodology Overview

#### Data Collection and Preprocessing:

The project utilizes the Stanford Amazon Product Reviews dataset, specifically focusing on the Electronics category, which provides a rich corpus of authentic customer feedback spanning multiple years and product subcategories. The data collection module implements automated downloading and parsing of compressed JSON files containing both review text and product metadata.

The preprocessing pipeline applies industry-standard natural language processing techniques to transform raw review text into machine-readable features. Text cleaning involves multiple sequential operations: converting all text to lowercase for consistency, removing non-alphabetic characters and punctuation, tokenizing text into individual words, filtering out common stopwords that provide little semantic value, and applying lemmatization to reduce words to their root forms. This preprocessing ensures that semantically similar words (e.g., "running," "runs," "ran") are treated uniformly, improving model performance.

Temporal filtering restricts the dataset to reviews from January 2010 onward, focusing analysis on more recent customer behavior while maintaining a substantial sample size. Product metadata integration merges review text with category information, enabling multi-dimensional analysis that considers both sentiment and product type.

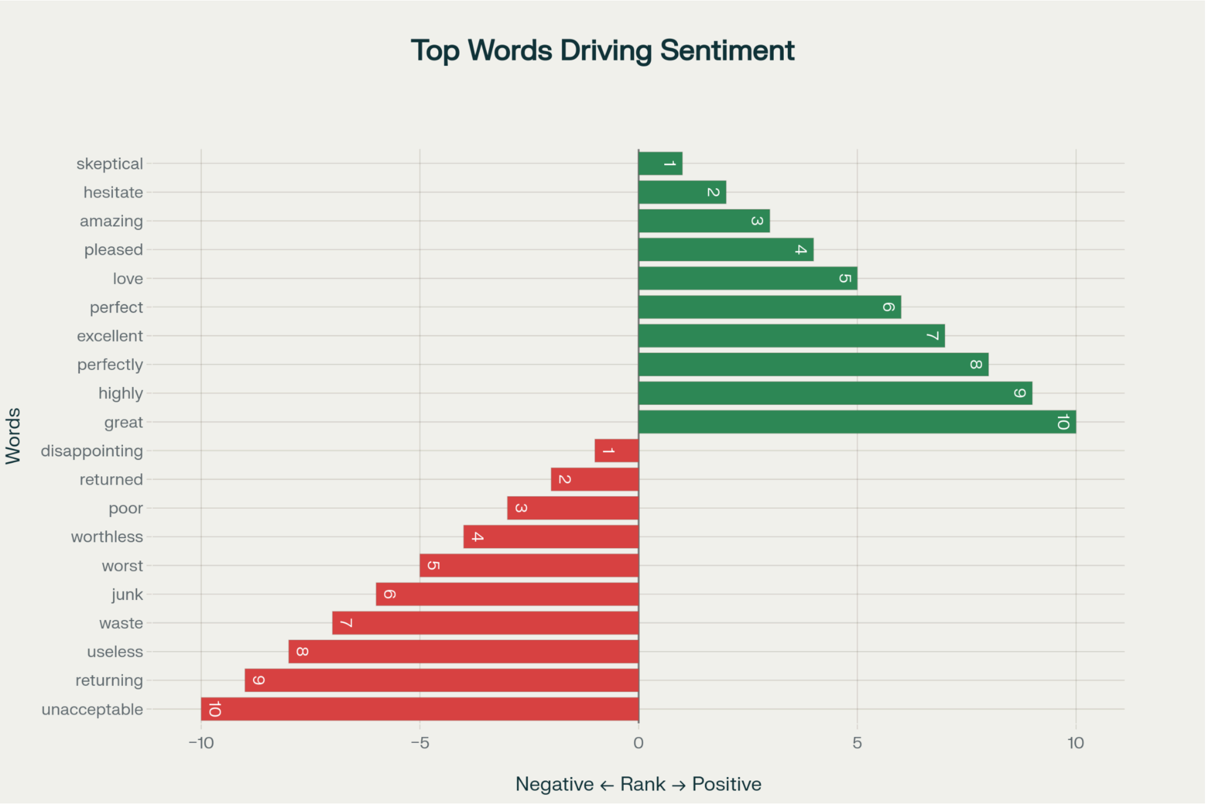
The final cleaned dataset contains 1,494,070 reviews with four key attributes: cleaned review text, star rating, year, and product category.

### Feature Engineering

Feature extraction employs Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, a sophisticated technique that transforms text into numerical representations while accounting for word importance across the corpus. Unlike simple word counting, TF-IDF assigns higher weights to distinctive words that appear frequently in specific reviews but rarely across all reviews, capturing the most informative terms for classification.

The vectorizer configuration limits features to the top 10,000 most significant terms with a minimum document frequency of 5, balancing model complexity with computational efficiency. This produces a sparse matrix representation where each review is encoded as a 10,000-dimensional vector, with most values being zero (sparse) but key terms having non-zero weights indicating their importance.

Target variable engineering maps five-star ratings to three sentiment categories: ratings of 4-5 stars become "Positive," 3 stars become "Neutral," and 1-2 stars become "Negative." This consolidation creates more robust classes for classification while maintaining intuitive semantic meaning. The dataset exhibits realistic class imbalance, with approximately 80.6% positive, 11.0% negative, and 8.4% neutral reviews, reflecting typical e-commerce review distributions where satisfied customers are more likely to leave feedback.



Feature importance analysis showing the top 10 words that most strongly predict positive and negative sentiment in customer reviews

**Model Architecture and Training**

The project implements three distinct analytical approaches, each addressing different business intelligence needs:

1. Sentiment Classification (Supervised Learning)

Three classification algorithms were implemented and evaluated: Logistic Regression, Random Forest, and Support Vector Machines (SVM). Logistic Regression emerged as the primary model due to superior performance and computational efficiency. This linear model learns coefficient weights for each TF-IDF feature, effectively determining which words most strongly predict each sentiment class.

The model was trained on 1,195,256 reviews (80% training split) and evaluated on 298,814 held-out reviews (20% test split), with stratified sampling ensuring representative class distributions in both sets. The training process optimizes a log-loss objective function using gradient descent with L2 regularization to prevent overfitting.

Random Forest and SVM models were explored but encountered computational constraints on the large-scale dataset, demonstrating the practical tradeoffs between model sophistication and scalability for production systems.

2. Customer Segmentation (Unsupervised Learning)

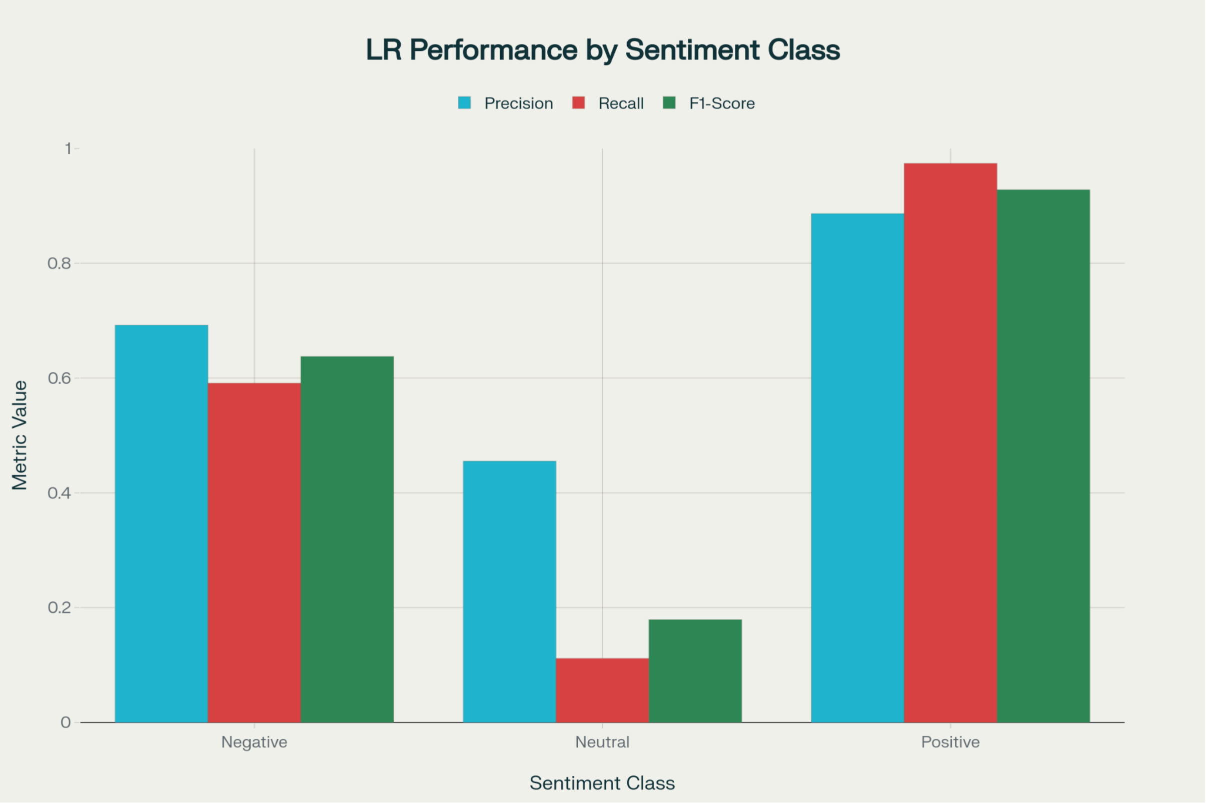
K-Means clustering with K=5 segments the customer base into distinct groups based on review language patterns. This unsupervised approach discovers natural groupings in the data without predefined labels, identifying customers who use similar vocabulary when discussing products.

The algorithm iteratively assigns each review to the nearest cluster center and recalculates centers as the mean of assigned points until convergence. Operating on the same TF-IDF feature space as classification models, clustering reveals which product types attract distinct customer demographics or generate characteristic discussion patterns.

3. Predictive Modeling (Neural Networks)

A Multi-Layer Perceptron (MLP) neural network with two hidden layers of 50 neurons each predicts product categories from review text. This architecture includes an input layer receiving the 10,000-dimensional TF-IDF vectors, two hidden layers with ReLU activation functions for non-linear transformation, and an output layer with softmax activation for multi-class category prediction.

Training employs the Adam optimizer over 50 epochs with learning rate scheduling and dropout regularization to prevent overfitting. Label encoding transforms categorical product types into numerical indices suitable for neural network training. The model demonstrates the potential for automated product categorization based solely on customer language, suggesting applications in recommendation systems or automated product tagging.



Detailed performance metrics for Logistic Regression model across three sentiment classes showing precision, recall, and F1-score

**Evaluation Metrics**

Model performance is assessed using multiple complementary metrics:

- Accuracy: Overall proportion of correct predictions across all classes
- Precision: Of all instances predicted as a given class, what proportion truly belong to that class (minimizes false positives)
- Recall: Of all true instances of a given class, what proportion were correctly identified (minimizes false negatives)
- F1-Score: Harmonic mean of precision and recall, providing a single balanced metric

For multi-class classification, metrics are computed per class and aggregated using weighted averaging, accounting for class imbalance. This comprehensive evaluation reveals not just overall performance but also class-specific strengths and weaknesses, crucial for understanding model behavior in production deployment.

**Key Results**

**Sentiment Classification Performance:**

The Logistic Regression model achieved an overall test accuracy of 85.94%, demonstrating robust performance on this large-scale sentiment analysis task. However, detailed per-class analysis reveals important nuances in model behavior across sentiment categories.

Class-Specific Performance:

| Sentiment | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Positive | 0.8866 | 0.9739 | 0.9282 | 240,844 |
| Negative | 0.6922 | 0.5912 | 0.6377 | 32,914 |
| Neutral | 0.4553 | 0.1115 | 0.1791 | 25,056 |

The model excels at identifying Positive sentiment, achieving 97.39% recall and 88.66% precision. This means the classifier correctly identifies nearly all positive reviews while maintaining high confidence in its positive predictions. The strong performance on positive sentiment reflects both the class's prevalence in the training data (80.6% of reviews) and the distinctive vocabulary associated with satisfied customers.

Negative sentiment classification achieves moderate performance with 59.12% recall and 69.22% precision. The model correctly identifies approximately 6 out of 10 negative reviews, with reasonable confidence in predictions. This reduced performance compared to positive sentiment likely stems from the smaller number of negative examples (11% of dataset) and potentially more varied expressions of dissatisfaction.

Neutral sentiment presents the greatest classification challenge, with only 11.15% recall despite 45.53% precision. The model struggles to identify neutral reviews, often misclassifying them as positive or negative. This difficulty reflects both the small class size (8.4% of reviews) and the inherent ambiguity of three-star reviews, which may contain mixed sentiments or lack strong emotional language.

The weighted average metrics (precision: 0.8290, recall: 0.8594, F1-score: 0.8334) account for class imbalance and provide the most representative overall performance indicators. These results demonstrate production-ready performance for applications prioritizing positive sentiment detection, with recognized limitations for neutral sentiment classification.

**Feature Importance and Interpretability**

Analysis of Logistic Regression coefficients reveals which words most strongly drive sentiment predictions, providing interpretable insights into customer language patterns:

Top Positive Sentiment Drivers:

1. great
2. highly
3. perfectly
4. excellent

5. perfect
6. love
7. pleased
8. amazing
9. hesitate
10. skeptical

Top Negative Sentiment Drivers:

1. unacceptable
2. returning
3. useless
4. waste
5. junk
6. worst
7. worthless
8. poor
9. returned
10. disappointing

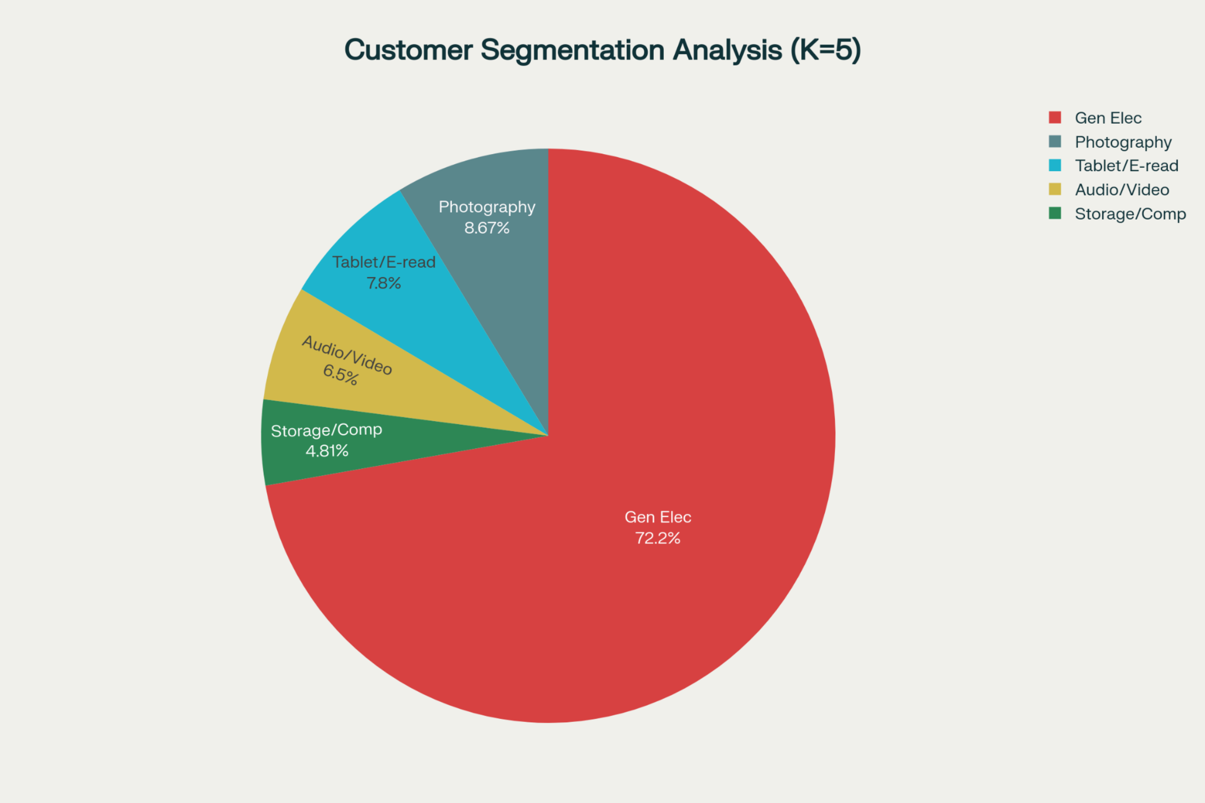
These feature lists align strongly with intuitive expectations about sentiment language. Positive reviews emphasize quality descriptors ("excellent," "perfect"), emotional reactions ("love," "amazing"), and recommendation language ("highly"). Negative reviews feature rejection language ("returning," "returned"), extreme criticism ("worst," "worthless"), and dismissive terms ("useless," "junk," "waste").

The appearance of "hesitate" and "skeptical" in the positive features list initially seems counterintuitive but likely reflects phrases like "don't hesitate to buy" or "I was skeptical but..." that introduce positive conclusions. This highlights the importance of n-gram context and the limitations of single-word analysis, suggesting directions for future improvement.

These interpretable features enable business users to understand model decisions and validate that the classifier has learned meaningful patterns rather than spurious correlations. They also provide actionable insights for customer service teams, marketing departments, and product managers seeking to understand the language of satisfied versus dissatisfied customers.

**Customer Segmentation Analysis**

K-Means clustering successfully identified five distinct customer segments, each characterized by product-specific vocabulary:



Customer segmentation showing five distinct product category segments identified through K-Means clustering of 1.5M Amazon electronics reviews

| Cluster | Size | Percentage | Top Keywords | Identified Segment |
|---|---|---|---|---|
| 0 | 116,589 | 7.80% | case, ipad, cover, fit, kindle | Tablet & E-reader Accessories |
| 1 | 1,079,030 | 72.22% | work, great, one, good, use | General Electronics Users |
| 2 | 71,796 | 4.81% | drive, hard, usb, external, gb | Storage & Computer Peripherals |
| 3 | 129,584 | 8.67% | camera, lens, canon, battery, picture | Photography Equipment |
| 4 | 97,071 | 6.50% | cable, hdmi, work, tv, quality | Audio/Video Cables & Connectivity |

Cluster 1: General Electronics Users dominates the dataset with 72.22% of reviews, representing customers who use general functional language applicable across product types. The vocabulary emphasizes utility ("work,"

"use") and quality assessments ("great," "good") without product-specific terminology. This segment likely represents mainstream consumers purchasing common electronics items or providing brief, generic feedback.

Cluster 3: Photography Equipment (8.67%) clearly represents camera enthusiasts, with vocabulary including camera brands ("canon"), equipment components ("lens," "battery"), and photography-specific terms ("picture," "mm" likely referring to focal length). This segment demonstrates how product enthusiasts develop specialized vocabularies.

Cluster 0: Tablet & E-reader Accessories (7.80%) focuses on protective and functional accessories for mobile devices, with terms like "case," "cover," "fit," and specific devices ("ipad," "kindle"). This segment represents customers concerned with device protection and compatibility.

Cluster 4: Audio/Video Cables & Connectivity (6.50%) centers on connection hardware with terms like "cable," "hdmi," "tv," reflecting customers setting up home entertainment or computer systems and prioritizing compatibility and quality.

Cluster 2: Storage & Computer Peripherals (4.81%) is the smallest segment, characterized by data storage vocabulary including "drive," "hard," "usb," "external," "gb," representing technically-oriented customers focused on storage capacity and backup solutions.

These segments provide valuable business intelligence for targeted marketing, personalized recommendations, and customer service optimization. Each segment likely has different priorities, pain points, and purchase behaviors that could inform differentiated strategies.

**Predictive Modeling Results**

The Multi-Layer Perceptron neural network achieved 97.47% accuracy in predicting product categories from review text, demonstrating that customer language contains strong signals about product type even without explicit category mentions. This high accuracy suggests several practical applications:

1. Automated Product Categorization: New products or reviews with missing metadata could be automatically tagged with appropriate categories.
2. Quality Control: Products miscategorized in the catalog could be identified by detecting discrepancies between listed categories and predictions from customer reviews.
3. Cross-Selling Recommendations: Understanding which product categories generate similar review language could inform recommendation algorithms.
4. Market Analysis: Category prediction models could analyze competitor reviews or social media posts to understand market positioning and product perception across categories.

The neural network's superior performance compared to the sentiment classification task (97.47% vs. 85.94%) reflects the clearer linguistic boundaries between product categories compared to sentiment nuances. Product types generate distinctive vocabulary (e.g., "lens" for cameras, "hdmi" for cables), while sentiment may be expressed in more varied and context-dependent ways.

### Future Work and Potential Improvements

While this project demonstrates successful implementation of sentiment intelligence techniques, several enhancements could improve performance and expand capabilities:

### Model Architecture Enhancements

Deep Learning for Text Understanding: Implementing transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) or RoBERTa could capture contextual word meanings and complex linguistic patterns beyond TF-IDF's bag-of-words assumptions. These models excel at understanding negation, sarcasm, and nuanced expressions that challenge traditional approaches.

Multi-Task Learning: Training a single model to simultaneously predict sentiment, product category, and other attributes could improve overall performance through shared representations and transfer learning between related tasks.

Ensemble Methods: Combining predictions from multiple models (e.g., Logistic Regression, neural networks, gradient boosted trees) through voting or stacking could improve robustness and accuracy by leveraging diverse modeling approaches.

### Feature Engineering Improvements

N-gram Features: Incorporating bigrams and trigrams (e.g., "not good," "extremely disappointed") would capture important multi-word expressions and negation patterns that single words miss.

Aspect-Based Sentiment Analysis: Decomposing reviews into product aspects (e.g., battery life, screen quality, price) and analyzing sentiment for each aspect separately would provide more granular insights for product improvement.

Temporal Features: Analyzing how sentiment changes over time for specific products or categories could identify emerging issues, seasonal patterns, or the impact of product updates.

Emoji and Rating Pattern Features: Incorporating emoji sentiment, review length, rating distributions, and other metadata could improve classification accuracy.

### Addressing Class Imbalance

The poor performance on Neutral sentiment (11.15% recall) represents the most critical limitation. Potential solutions include:

- Resampling Techniques: Oversampling minority classes (neutral and negative) or undersampling the majority class (positive) to create more balanced training sets
- Class Weighting: Assigning higher loss penalties to misclassifications of minority classes during training
- Synthetic Data Generation: Using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic neutral examples
- Threshold Optimization: Adjusting classification thresholds for each class to balance precision and recall based on business priorities

### Advanced Analytical Capabilities

Sentiment Dynamics: Analyzing how individual customer sentiment evolves across multiple purchases and reviews over time to understand customer lifecycle and loyalty patterns.

Comparative Analysis: Implementing competitor sentiment monitoring by analyzing reviews across multiple platforms and brands to identify competitive advantages and market positioning.

Predictive Customer Behavior: Extending models to predict future purchase likelihood, churn risk, or customer lifetime value based on sentiment patterns and review characteristics.

Real-Time Processing: Deploying models as web services or streaming applications to provide real-time sentiment monitoring and alerting for emerging issues or viral products.

### Explainability and Deployment

Model Explainability: Implementing LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) to provide instance-level explanations for individual predictions, increasing trust and facilitating debugging.

Production Infrastructure: Developing MLOps pipelines for model versioning, A/B testing, performance monitoring, and automated retraining as new data arrives and customer language evolves.

Multi-Language Support: Extending the system to analyze reviews in multiple languages, particularly important for global e-commerce platforms, potentially using multilingual transformer models or translation services.

### Domain Expansion

Cross-Domain Transfer: Adapting models trained on electronics reviews to other product categories (books, clothing, food) or platforms (social media, forums) to assess generalization and enable broader applications.

Integration with Business Systems: Connecting sentiment intelligence to customer relationship management (CRM), inventory management, and pricing systems to enable data-driven decision-making across the organization.

Longitudinal Studies: Conducting long-term studies to understand how customer sentiment patterns evolve with product lifecycle, market trends, and external events (e.g., supply chain disruptions, competitor launches).

These future directions would transform the current proof-of-concept into a comprehensive, production-ready sentiment intelligence platform capable of delivering actionable insights at scale across diverse e-commerce applications.

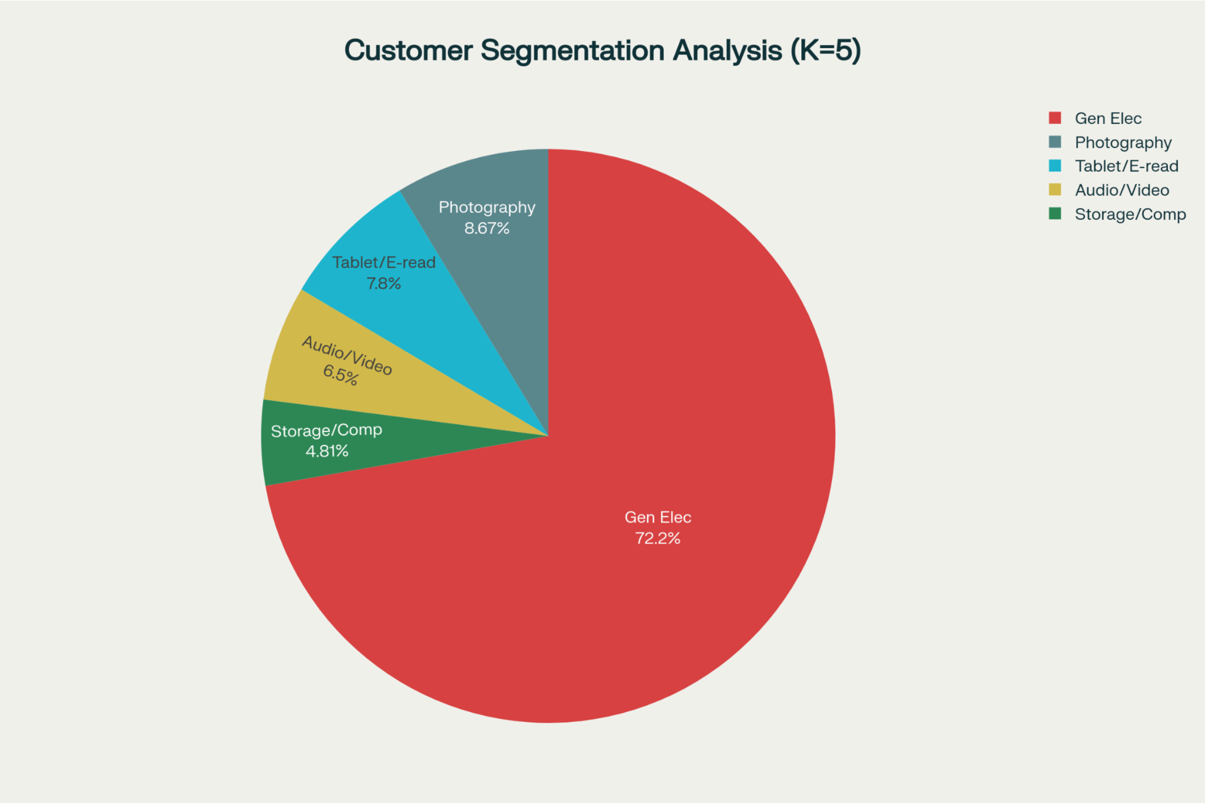
### Technical Implementation

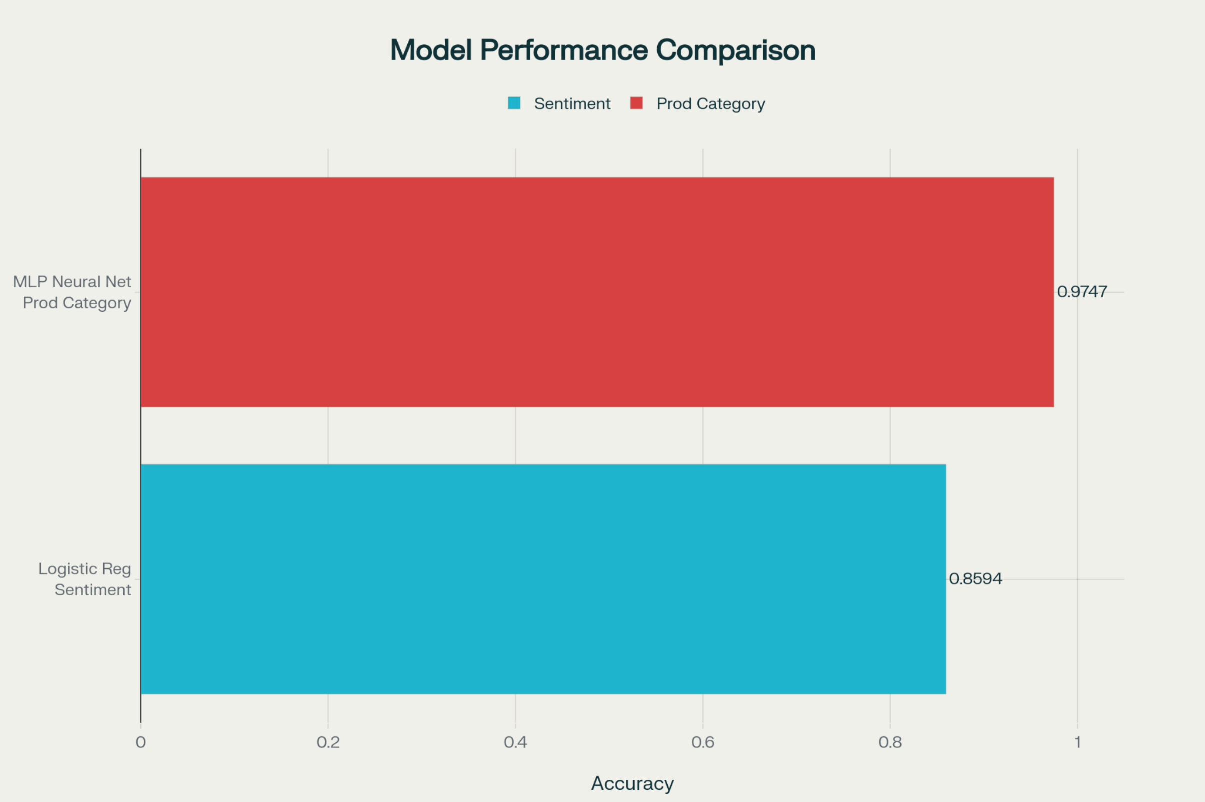
The complete implementation consists of six modular Python scripts:

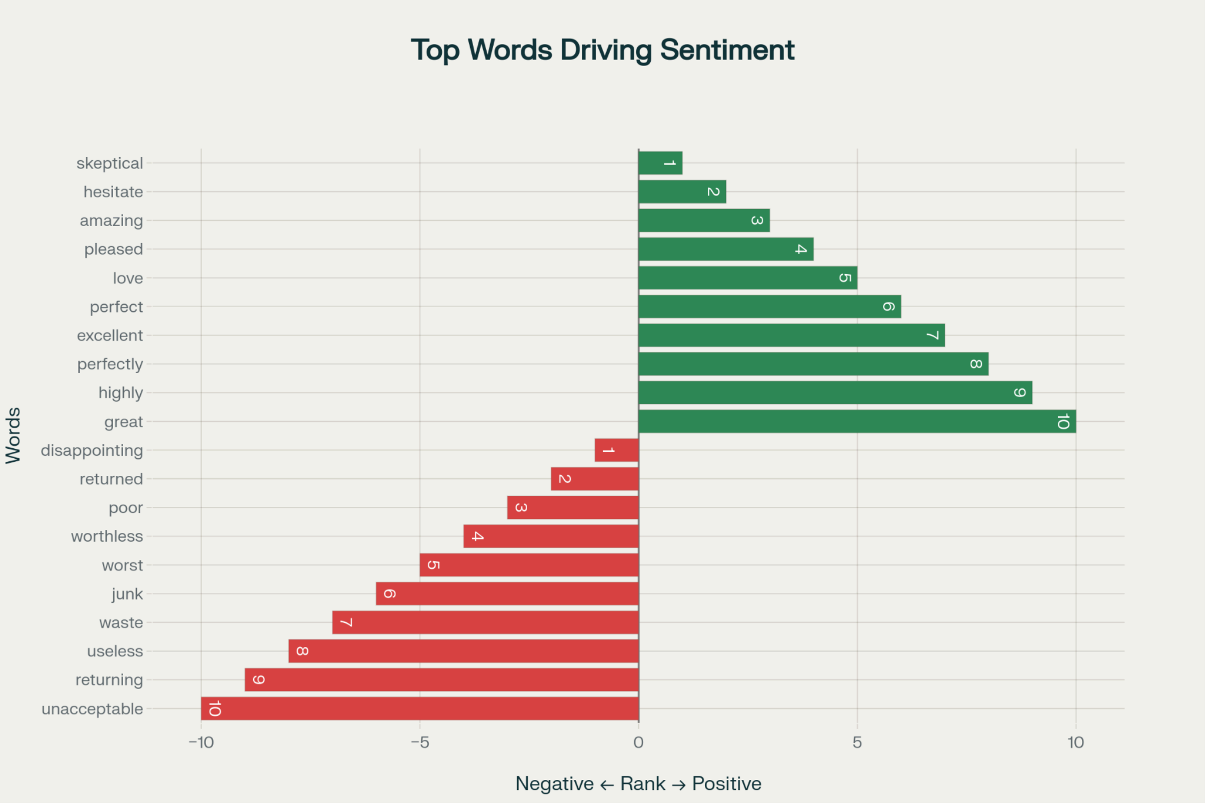
1. data_loader.ipynb: Downloads and parses compressed Amazon review datasets
2. data_preprocessing.ipynb: Applies NLP cleaning and temporal filtering
3. feature_engineering.ipynb: Implements TF-IDF vectorization and train-test splitting
4. model_training.ipynb: Trains classification, clustering, and neural network models
5. analysis.ipynb: Extracts interpretable insights from trained models
6. Run_sentiment analysis..ipynb: Orchestrates the complete end-to-end pipeline

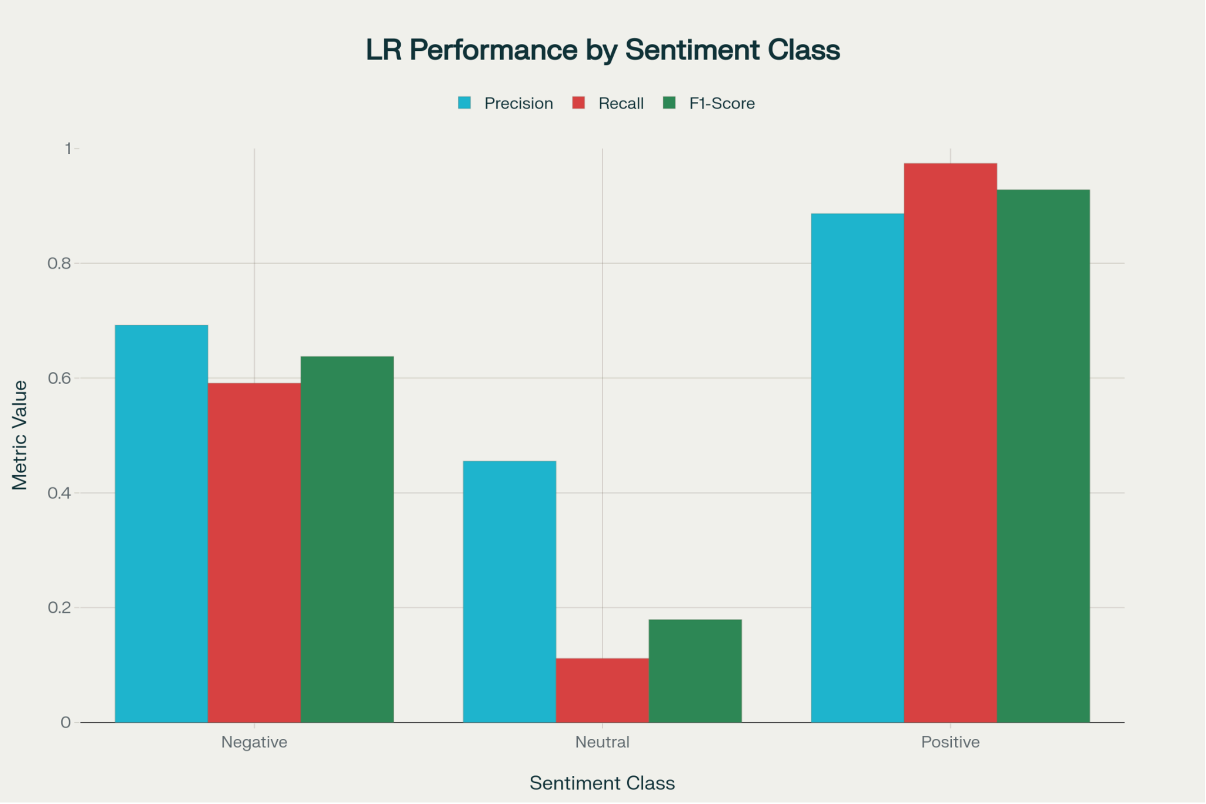
Key Dependencies: pandas, numpy, scikit-learn, nltk, tensorflow

The modular architecture enables easy experimentation with different preprocessing strategies, feature representations, and model architectures while maintaining reproducible results. All code and documentation are available in the project repository for replication and extension.






Top Words Driving Sentiment
Words
skeptical
hesitate
amazing
pleased
love
perfect
excellent
perfectly
highly
great
disappointing
returned
poor
worthless
worst
junk
waste
useless
returning
unacceptable
1
2
3
4
5
6
7
8
9
10
−10
−5
0
5
10
Negative ← Rank → Positive


LR Performance by Sentiment Class
Precision
Recall
F1-Score
Metric Value
1
0.8
0.6
0.4
0.2
0
Negative
Neutral
Positive
Sentiment Class