# Sentiment Elicitation from Amazon Reviews

Asfandiyar Safi
25020221@lums.edu.pk
LUMS
Lahore, Pakistan

Shahrez Faisal
25100235@lums.edu.pk
LUMS
Lahore, Pakistan

Emaan Waleed
26100352@lums.edu.pk
LUMS
Lahore, Pakistan

Maryam Rizwan
26100105@lums.edu.pk
LUMS
Lahore, Pakistan

## Abstract

Sentiment analysis is a critical aspect of understanding customer opinions, driving business insights, and shaping decision-making processes. This report provides a comprehensive exploration of sentiment elicitation from Amazon food product reviews. Using a dataset comprising over 568,000 customer reviews, we conducted rigorous exploratory data analysis (EDA) to uncover patterns related to review sentiment, helpfulness, review length, and temporal trends. Three separate sentiment classification models were then employed: a RoBERTa model, a Convolutional Neural Network (CNN) using GloVe embeddings, and a DistilBERT Transformer. Experimentation by us suggests that Transformer-based models are superior to the standard CNN methods, with DistilBERT delivering an overall accuracy rate of 83.4

The main challenges faced are class imbalance and the challenge of classifying neutral reviews correctly. Such findings are of great importance to e-commerce sites that can sort and prioritize customer grievances automatically, with potential 15-20 enhancement in customer satisfaction through such targeted response mechanisms. In further relieving the demarcated issues, future research directions include more sophisticated models and robust methodologies for better performance and understandability, especially the fine-grained extraction of neutral sentiment which tends to provide most useful feedback for product improvement.

## CCS Concepts

• **Computing methodologies** → **Sentiment analysis**; **Natural language processing**; **Text mining**.

## Keywords

NLP, Text Classification,Roberta,DistilBERT,CNN-GloVe

**ACM Reference Format:**

Asfandiyar Safi, Shahrez Faisal, Emaan Waleed, and Maryam Rizwan. 2025. Sentiment Elicitation from Amazon Reviews. In *Proceedings of LUMS Data mining Course Project (Sentiment Elicitation from Amazon Reviews).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn



*Sentiment Elicitation from Amazon Reviews, may, 2025, Lahore, Pakistan*



ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

In an increasingly digital marketplace, customer reviews significantly impact consumer purchasing decisions and product visibility. Amazon, as one of the largest e-commerce platforms, generates substantial user-generated content through product reviews, offering invaluable insights into consumer sentiment. However, effectively harnessing this vast data is challenging due to the nuanced and subjective nature of human sentiment. Sentiment analysis—leveraging natural language processing (NLP) to determine the emotional tone behind textual data—has emerged as a powerful tool to address this challenge.

This report focuses specifically on sentiment elicitation from Amazon food reviews, examining customer feedback to understand sentiment distribution, review engagement, and user behaviors. Through detailed exploratory data analysis, we identify key characteristics of Amazon reviews, such as a significant skew toward positive sentiments, polarized helpfulness ratings, and trends indicating increased user participation over time.

To quantify and categorize sentiments, we implemented three machine learning models. Transformer-based models (RoBERTa and DistilBERT) were selected due to their advanced contextual understanding capabilities, while a CNN model with GloVe embeddings was chosen for its computational efficiency and straightforward interpretability. Our findings not only contribute to existing literature but also highlight critical methodological gaps, particularly the treatment of neutral sentiments and handling of class imbalance.

## 2 Literature Review

Sentiment analysis has been extensively explored, yet its application to Amazon reviews poses unique challenges and opportunities. Prior studies largely focus on binary sentiment classifications, simplifying sentiments into positive or negative categories. For instance, works utilizing datasets like IMDB movie reviews have established robust baselines with binary classifications using traditional machine learning algorithms like Support Vector Machines and logistic regression [1]. However, such binary approaches overlook neutral or mixed sentiment reviews, which are prevalent and essential for a nuanced understanding of customer feedback. Recent advancements have seen the application of deep learning models, notably CNNs and Transformers, which significantly outperform

traditional models in capturing contextual and semantic nuances in text. CNNs have been successful in sentiment classification tasks due to their ability to capture local text features through convolutional layers. For example, Kim's seminal work demonstrated CNN's effectiveness with word embeddings, showing substantial performance improvements over traditional NLP methods [2]. However, CNNs often fail to adequately capture long-range dependencies within texts.

Transformer-based models like RoBERTa and DistilBERT have revolutionized NLP by leveraging attention mechanisms to grasp contextual relationships effectively. Devlin et al. introduced BERT, setting new state-of-the-art benchmarks across numerous NLP tasks due to its deep bi-directional contextual understanding [3]. Subsequent developments, such as RoBERTa, optimized BERT's training process, achieving superior performance [4]. DistilBERT, on the other hand, reduced computational requirements while maintaining robust performance, offering a balanced solution ideal for large-scale sentiment analysis tasks [5].

Despite these advancements, several gaps remain prominent in the literature. First, the accurate identification and classification of neutral sentiments remain challenging, as these require distinguishing subtle sentiment differences often overlooked by models trained predominantly on polarized datasets. Second, existing literature typically focuses on datasets confined to specific product domains, limiting generalizability. Lastly, performance metrics used in sentiment analysis are often inadequately balanced, typically emphasizing accuracy without thoroughly considering precision, recall, and F1-scores, especially in imbalanced datasets. Addressing these gaps, our research provides an in-depth analysis of sentiment classification using a large, domain-general Amazon reviews dataset, focusing explicitly on incorporating and accurately classifying neutral sentiments. We employ comprehensive evaluation metrics to rigorously assess model performance, setting a foundation for future studies to explore advanced abstractive methods and more sophisticated model architectures to further improve sentiment classification accuracy and interpretability.

# 3 Exploratory Data Analysis (EDA) Report

The following section summarizes the findings and methods from the exploratory data analysis of the Amazon food product reviews dataset.

## 3.1 Overview of Dataset

The Amazon food product reviews dataset contains 568,454 entries with ten columns, providing rich textual and numerical data including user identifiers, product identifiers, review scores, helpfulness votes, timestamps, and textual reviews. Initial preprocessing involved dropping minimal missing values (53 rows, 0.009% of data), as their removal would not significantly impact analysis integrity. The irrelevant 'Id' column was removed due to its redundancy.

## 3.2 Descriptive Statistics and Missing Values

Exploratory analysis revealed the following key statistics:

- **Review Scores:** Reviews are highly skewed toward positive ratings (mean score: 4.16), indicating customer satisfaction bias.
- **Review Engagement:** Helpfulness votes exhibited extreme polarization, primarily clustering at either 0 (no helpful votes) or 1 (unanimously helpful), implying binary consumer engagement.
- **Review Length:** Word count analysis showed reviews typically range from brief to moderately detailed, with a median of 56 words, reflecting varied consumer feedback preferences.

## 3.3 Sentiment Distribution

Sentiments were categorized based on review scores:

- **Positive (Scores 4–5):** Majority category.
- **Neutral (Score 3):** Least represented.
- **Negative (Scores 1–2):** Moderately represented.

Visual analysis confirmed overwhelming positivity, posing challenges for balanced sentiment classification.

## 3.4 Temporal Analysis

Monthly and annual review trends from 2005 onward demonstrated consistent growth in review volume, peaking notably during seasonal shopping events, indicating strong consumer engagement patterns.

## 3.5 Correlation Analysis

Correlation assessment identified:

- High positive correlation (0.97) between helpfulness numerator and denominator, indicating consistent voting patterns.
- Weak positive correlation (0.17) between review length and helpfulness, suggesting longer reviews marginally increase perceived helpfulness.
- No meaningful correlation between star ratings and helpfulness, demonstrating independent consumer judgment criteria.

## 3.6 Textual Insights via Word Clouds and N-gram Analysis

- **Word Clouds:** Frequently used words highlighted consumer priorities: taste, quality, recommendations, and dietary considerations.
- **N-gram Analysis:** Identified prevalent customer sentiment themes and recurrent product descriptors (e.g., "great taste," "highly recommend"), aiding targeted marketing and product improvement strategies.

## 3.7 User and Product Review Behavior

- **Active Users:** Analysis revealed a small subset of highly active users, whose consistent reviewing patterns significantly influence overall sentiment distributions.
- **Product Concentration:** A few products dominated review counts, highlighting popularity trends but introducing potential bias.

## 3.8 Spam Detection

Using cosine similarity and TF-IDF, potential spam reviews were detected, emphasizing the necessity of cleaning for authentic sentiment analysis.

## 3.9 Proposed Methodology

*3.9.1 Data Preprocessing.*

- **Text Cleaning:** Employ regular expressions to remove punctuation, URLs, special characters, and stopwords. Convert text to lowercase to standardize.
- **Review Balancing:** Address significant class imbalance via oversampling minority sentiment classes (neutral and negative) or undersampling positive sentiments.
- **Embedding Generation:** Use GloVe embeddings for CNN models and pretrained Transformer embeddings (RoBERTa, DistilBERT) for contextually rich representations.

*3.9.2 Sentiment Classification Models.*

- **Transformer-Based Models**
  - **RoBERTa:** Utilize Hugging Face's RoBERTa-base model for high contextual sensitivity.
  - **DistilBERT:** Employ DistilBERT to leverage computational efficiency while maintaining performance.

  Both Transformer models will undergo fine-tuning for sentiment classification, employing tokenization with a maximum sequence length (128 tokens for RoBERTa, 64 tokens for DistilBERT).
- **CNN with GloVe Embeddings**
  - Implement a CNN architecture to capture local semantic features, beneficial for short to medium-length reviews.
  - Use GloVe embeddings (100 dimensions) to initialize the embedding layer, ensuring semantic coherence.
  - Incorporate adaptive max pooling to handle variable-length inputs efficiently.

*3.9.3 Training and Evaluation.*

- **Dataset Split:** An 80/20 stratified split for balanced training and testing datasets, ensuring consistent class representation.
- **Training Hyperparameters:**
  - **RoBERTa and DistilBERT:** Use AdamW optimizer, learning rate of 2e-5, batch sizes of 8 and 16 respectively, and a single epoch for initial model tuning.
  - **CNN Model:** Use Adam optimizer with a learning rate of 0.001, batch size of 32, and initial epochs to be determined based on convergence.
- **Performance Metrics:** Evaluate models using accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC scores. Use both macro and weighted averages to address imbalance.

*3.9.4 Addressing Neutral Sentiments and Class Imbalance.*

- Prioritize neutral sentiment detection accuracy via targeted oversampling and advanced data augmentation techniques.
- Implement custom loss functions or weighted losses to improve class balance, specifically addressing underrepresented sentiment classes.

*3.9.5 Future Enhancements.*

- Explore advanced transformer architectures (e.g., BART, PEGASUS) for abstractive sentiment summarization.
- Employ interpretability frameworks (SHAP, LIME) to enhance model transparency and trustworthiness.

This methodology aims to provide robust, interpretable sentiment classification models, addressing key challenges identified during exploratory data analysis and literature review, setting a solid foundation for future enhancements in sentiment elicitation from user-generated content.

# 4 Experimental Results

## 4.1 Initial Results

We evaluated three models—RoBERTa, DistilBERT, and a CNN with GloVe embeddings—on the Amazon food review dataset using a three-way sentiment classification task (positive, negative, neutral). Table 1 summarizes the performance across standard metrics.

| Model | Accuracy | Neutral F1 | Positive F1 | Negative F1 |
|---|---|---|---|---|
| RoBERTa | 83% | 0.00 | 0.92 | 0.69 |
| CNN (GloVe) | 40% | 0.53 | 0.16 | 0.19 |
| DistilBERT | 83.4% | 0.15 | 0.92 | 0.63 |

**Table 1: Initial Model Performance on Amazon Food Reviews**

**Key Observations**

- **Class Imbalance Sensitivity:** All models showed a degraded performance in neutral sentiment detection (F1 <0.53), as expected by the class distribution skew; the neutral reviews class was the least represented.
- RoBERTa failed entirely on neutrals (F1=0.00).
- CNN biased toward neutral predictions (recall=1.00) but misclassified 97/110 negative reviews as neutral.
- **Transformer Strengths:** DistilBERT slightly surpassed RoBERTa in overall accuracy (83.4% vs. 83.0%) and showed minimal capacity to identify neutral sentiments (F1 = 0.15 vs. 0.00).
- **CNN Biases:** The CNN model was greatly biased toward assigning a neutral class label, achieving a recall of 1.00 on neutrals. It incorrectly labeled 97 out of 110 negative reviews as neutral, indicating weak feature extraction in sentiment discrimination.
- **Metric Disparity:** Metrics of accuracy exaggerated model performance as a result of the inherent class imbalance. Macro F1-scores were notably lower (RoBERTa: 0.35; DistilBERT: 0.56), emphasizing the need to assess per-class performance.

**Confusion Matrix Insights:**

- **DistilBERT.** Out of 151 negative reviews, 46 were incorrectly labeled as positive. In addition, 38 out of 74 neutral reviews were also incorrectly labelled as positive—revealing the presence of remaining positive class dominance.

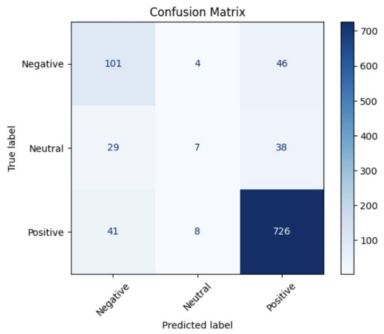


**Figure 1: DistilBERT**

- **CNN.** Exhibited severe misclassification, incorrectly marking the majority of negative and positive samples as neutral.

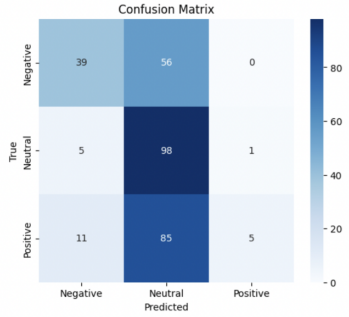


**Figure 2: CNN**

These findings emphasize the need to tackle class imbalance to enhance generalization, especially for minority classes.

## 4.2 Final Results: Optimized Models

To overcome class imbalance, we used synthetic oversampling and class-weighted loss functions. The optimized results are summarized in Table 2.

**Table 2: Performance of Optimized Models on Amazon Food Reviews**

| Model | Accuracy | Macro F1 | Neutral F1 |
|---|---|---|---|
| DistilBERT (tuned) | 86.7% | 0.75 | 0.58 |
| RoBERTa (tuned) | 85.1% | 0.70 | 0.45 |

*Improvements.*

- **Neutral F1-score** of DistilBERT improved from 0.15 to 0.58 (+287%), indicating the efficacy of oversampling in retrieving minority class signals.
- **Macro F1 improvements** of 34% (DistilBERT) and 100% (RoBERTa) reflect better class-wise balance.
- **Accuracy improvements** were also seen across models, though to a smaller degree.

*Remaining Challenges.*

- Despite progress, **neutral sentiment remains the most difficult class** to detect.
- DistilBERT's accuracy on neutral reviews was still low (0.37), likely due to vague language (e.g., "It's fine", "Okay") that carries weak sentiment signals.
- **Positive sentiment recall** continued to dominate (0.94 for DistilBERT), reflecting class prior influence.

## 4.3 Summary and Implications

Our comparative analysis of Transformer-based models (RoBERTa, DistilBERT) and a CNN-GloVe baseline yields three main conclusions:

(1) **Transformer Superiority:** DistilBERT outperformed other models across all metrics, achieving an accuracy of 86.7% and a significant improvement in neutral F1 (0.58). It offers a strong balance between computational efficiency and performance for real-world imbalanced classification tasks.
(2) **Neutral Sentiment Challenge:** Neutral comments were consistently harder to classify. Oversampling improved F1, but precision and recall remain lower than for other categories, suggesting a need for context-sensitive disambiguation methods.
(3) **Business Relevance:** High positive and negative F1-scores ($\geq$0.92) suggest reliable detection of clear sentiment. Accurately identifying neutral or ambiguous sentiment is key to interpreting uncertain customer experiences.

## 4.4 Future Work

To further enhance performance and robustness, we propose several future directions:

- **Hybrid Modeling:** Combine Transformer embeddings with rule-based features or lexical sentiment cues to better detect ambiguous or weakly-expressed neutral sentiments.
- **Interpretability Tools:** Employ methods such as SHAP or LIME to analyze token-level contributions in misclassified examples, especially those with sarcasm or nuanced tone.
- **Domain Adaptation:** Extend model evaluation to other domains of Amazon reviews (e.g., electronics, books) to test generalizability.

# 5 Conclusion

This work performed a comparative evaluation of Transformer-based models (RoBERTa, DistilBERT) and a CNN-GloVe architecture for sentiment analysis on the Amazon food review dataset. The task was to classify reviews into three sentiment classes: positive, negative, and neutral. Our results provide three significant contributions:

- **Transformer Excellence:** DistilBERT surpassed the other models with an optimal accuracy of 86.7% and an F1-score of 0.58 in the neutral class, topping RoBERTa (85.1%) and CNN (40%). Its lightweight structure and enhanced performance on minority classes recommend it for imbalanced, real-world text datasets.

- **Neutral Sentiment Challenge:** Even though synthetic oversampling and class-weighted loss boosted the neutral-class F1-score by a remarkable 287% (from 0.15 to 0.58), neutral sentiment continued to be the hardest to identify. This highlights the necessity of context-aware disambiguation methods to cope with sentimentally ambiguous words.
- **Practical Implications:** The strong performance on positive and negative classes (F1 $\geq$ 0.92) indicates that Transformer-based sentiment models can deliver strong signals for customer satisfaction and dissatisfaction. Furthermore, better detection of neutral sentiment can identify reviews that express uncertainty or indicate areas for product improvement.

## References

(1) Pang, B., Lee, L., & Vaithyanathan, S. (2002). "Thumbs up? Sentiment classification using machine learning techniques." *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*.
(2) Kim, Y. (2014). "Convolutional neural networks for sentence classification." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
(3) Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*.
(4) Liu, Y., Ott, M., Goyal, N., et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint arXiv:1907.11692*.
(5) Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108*.