Topics In Data Science Final — Amazon Reviews

Alex Bean
Cornell Tech - ORIE
atb95@cornell.edu

Utkarsh Goyal
Cornell Tech - ORIE
ug42@cornell.edu

Rishab Jain
Cornell Tech - ORIE
rj424@cornell.edu

Abstract

This study leverages sentiment analysis of Amazon reviews to uncover actionable insights that inform product development and brand management strategies under the Prediction, Computation, and Stability (PCS) framework. Using a balanced training dataset of 1.8 million reviews and a complementary testing dataset, the analysis focuses on identifying influential words and recurring complaints, particularly in negative reviews, to pinpoint product-specific issues and prioritize areas for improvement. A robust preprocessing pipeline, including data cleaning, text standardization, and TF-IDF vectorization, supported reliable sentiment classification through models such as Logistic Regression, XGBoost, and Long-Short Term Memory (LSTM) networks. Logistic Regression emerged as the most practical model balancing accuracy, efficiency, and robustness, while LSTM demonstrated superior predictive performance at higher computational costs. The findings highlight opportunities to refine product offerings by addressing recurring pain points, enhancing customer satisfaction, and revealing nuanced feedback beyond star ratings. Future directions include implementing topic modeling, tailoring analysis to product categories, and tracking long-term sentiment trends to further strengthen Amazon's brand perception.

1 Introduction

Customer sentiment plays a pivotal role in shaping brand perception and driving product development decisions. In today's competitive e-commerce landscape, understanding the nuances of customer feedback is essential for maintaining customer satisfaction and loyalty. Amazon's extensive review ecosystem provides an unparalleled opportunity to analyze sentiment at scale, uncovering actionable insights that inform strategic priorities. This paper addresses the critical question: *"How can the most influential words identified through binary sentiment classification of Amazon reviews guide strategies to improve Amazon's brand perception?"* By leveraging sentiment analysis, we aim to identify recurring themes in customer feedback, particularly negative sentiment, to pinpoint areas for targeted improvements and refine product offerings. This approach enables Amazon to address dissatisfaction proactively while strengthening brand loyalty.

The primary objective of this study is to explore how textual reviews can transcend traditional metrics, such as star ratings, to reveal nuanced insights into customer sentiment. Using a large, balanced training dataset of 1.8 million reviews, complemented by a richer but imbalanced testing dataset, the analysis focused on extracting key predictors of sentiment and their contextual significance. Negative sentiment reviews, in particular, were analyzed to uncover product-specific issues, such as deficiencies in functionality (software or hardware), that may not be evident in aggregated ratings. By identifying these recurring complaints and frequently mentioned features, this analysis offers practical recommendations for prioritizing product development efforts and addressing critical customer concerns.

To ensure the reliability and interpretability of our findings, the analysis was underpinned by a robust data preprocessing pipeline. This included steps such as removing URLs, special characters, and

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

redundant punctuation, as well as standardizing text through lowercase conversion, abbreviation expansion, and Porter Stemming. Additionally, advanced modeling techniques, including Logistic Regression and neural networks, were employed to classify sentiment and uncover influential words associated with dissatisfaction. By combining sentiment analysis with topic modeling and category-specific insights, this paper aims to provide a framework for understanding customer sentiment, addressing recurring issues, and tracking trends that inform long-term product strategies. The findings underscore the value of integrating advanced analytics with customer feedback to refine products and enhance Amazon's brand perception.

2 Overview of Data

Two distinct datasets were employed in this analysis, both sourced from Kaggle, offering complementary insights into Amazon review data. The first dataset originates from the Stanford Network Analysis Project (SNAP), which originally contained approximately 34.7 million reviews spanning 18 years up to March 2013. This dataset includes reviews from over 6.6 million users covering 2.4 million products, serving as a comprehensive repository of textual feedback. For the purposes of this study, a smaller but still substantial subset of the SNAP dataset was utilized. In this subset, reviews with ratings of 1 and 2 were labeled as negative, while ratings of 4 and 5 were labeled as positive. Reviews with a score of 3, considered neutral, were excluded due to their limited sentimental value. The dataset had a balanced distribution, comprising 1.8 million training samples and 200,000 validation samples for each sentiment, facilitating binary classification.

A secondary dataset, also sourced from Kaggle, was employed as the testing set. This dataset consists of over 34,000 consumer reviews of Amazon products, providing a more recent and nuanced view of customer feedback for the period between 2014 and 2017. It includes richer metadata such as product details (e.g., name, category, manufacturer) and review attributes (e.g., date added, star rating, purchase status, recommendation status, username). To ensure compatibility, this dataset underwent a transformation process to align with the format of the training and validation data. After standardization, the test set exhibited a more realistic and imbalanced sentiment distribution, containing approximately 800 negative reviews and 32,000 positive reviews (with the remaining neutral sentiment omitted).

This methodological approach provides an opportunity to evaluate model stability under the PCS (Prediction, Computation, and Stability) framework. Given that the testing data is more recent than the training and validation data, it allows an assessment of the model's consistency and robustness against natural variations, such as changes in language and writing styles. By analyzing the model's performance on this unbalanced and more realistic dataset, the findings derived from the primary dataset are tested for their generalizability, thereby enhancing confidence in their applicability to real-world scenarios.

3 Exploratory Data Analysis

3.1 Histogram

To better understand the structure of the Amazon review data, an analysis of review length distribution was conducted, segmented by sentiment polarity.

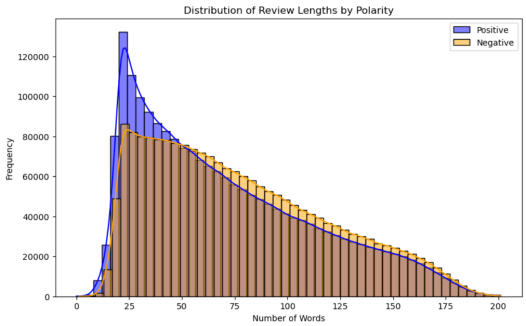


Figure 1: Distribution of Review Lengths by Polarity

As shown in Figure 1, the majority of reviews are relatively short, with a pronounced peak around 20 words. This indicates that customers tend to be concise when expressing their opinions. The distribution displays a long tail, with relatively few reviews exceeding 200 words. To simplify visualization, reviews longer than 200 words were grouped into a single bucket, accounting for approximately 700 reviews. The longest recorded review, spanning nearly 1,800 words, was positive, suggesting that customers are more likely to provide detailed feedback when they are satisfied with a product.

The histogram also reveals notable differences in review lengths between positive and negative sentiments. Positive reviews exhibit a sharper, more pronounced peak near 20 words, suggesting that satisfied customers often use brief and direct language. Conversely, negative reviews demonstrate greater variability, with a broader and flatter distribution extending into longer word counts. This suggests that dissatisfied customers are more inclined to elaborate on their grievances or provide contextual details.

The imbalance in review lengths has implications for the sentiment classification model. Shorter reviews may lack sufficient context, potentially leading to challenges in accurately predicting sentiment. On the other hand, longer reviews offer richer and more nuanced information, which can contribute to model robustness.

3.2 World Cloud

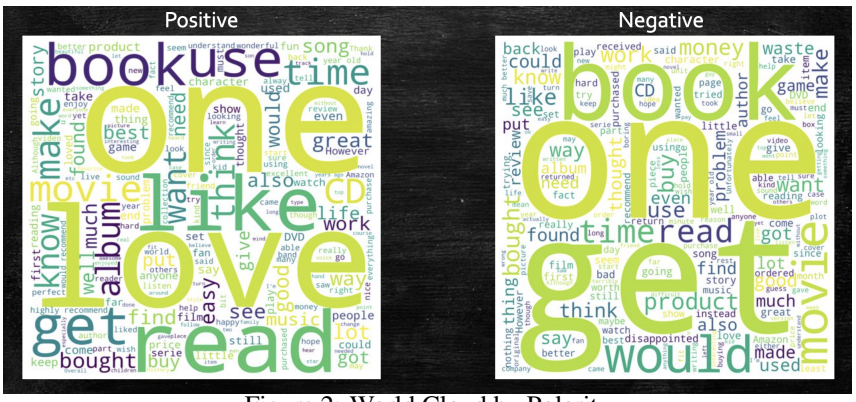


Figure 2: World Cloud by Polarity

Word clouds were generated for each sentiment category to visualize the most frequently used words in the dataset. These visualizations provide an intuitive summary of text data, with the size of each word reflecting its frequency. However, the word clouds revealed a significant limitation in the raw data: without text cleaning or preprocessing, the most frequent words, such as “book,” “one,” “read,” and “get,” dominated both positive and negative reviews, failing to convey sentiment-specific meanings. For example, terms like “book” and “read” in positive reviews do not inherently signify positive emotions, while words like “get” and “book” in negative reviews lack clear negative connotations.

This observation highlighted the insufficiency of raw word frequency as a standalone indicator for sentiment differentiation. Consequently, text preprocessing steps were implemented to refine the dataset, ensuring that the analysis would capture more meaningful and sentiment-relevant patterns.

4 Pre-processing

4.1 Data Cleaning

The data cleaning process was meticulously designed to enhance the quality and consistency of the textual data, ensuring that the dataset captured meaningful content while eliminating noise. Key steps included the removal of URLs, HTML entities, special characters, numbers, and extraneous punctuation. These elements, while common in raw textual data, do not contribute to understanding sentiment and were therefore excluded to streamline the dataset.

Standardization was another critical step. Abbreviations, such as replacing "u" with "you," were normalized to ensure uniformity across the dataset. Additionally, HTML tags, non-ASCII characters, and single quotes were removed to create a cleaner and more consistent representation of the text. Converting all text to lowercase further eliminated variations caused by capitalization, and punctuation was stripped to focus on the core semantic content. Stop words—common words like "the," "is," and "and"—which do not carry significant meaning in sentiment analysis, were also removed to reduce noise.

To further optimize the dataset, Porter Stemming was applied, a technique that reduces words to their root form. For example, words such as "running" and "runs" were converted to "run," while "happily" was reduced to "happili." This step helped minimize redundancy by consolidating variations of the same word, preserving their core meaning while reducing the dimensionality of the data. By focusing on root words, the preprocessing ensured that the models could better generalize patterns in the data.

To evaluate the effectiveness of the cleaning process, the word clouds were regenerated after preprocessing. The updated word clouds showed significant improvement, with sentiment-specific terms emerging more prominently, indicating that the dataset was now better equipped to support meaningful sentiment analysis. These refinements provided confidence that the data preparation steps were successfully aligning the dataset with the objectives of the study.

Figure 3: Cleaned Word Cloud by Polarity

The cleaned word clouds, shown in Figure 3, offer a clear and sentiment-specific visualization of the most frequent terms in positive and negative reviews following preprocessing. In positive reviews, words such as "recommend," "highly," "best," "works," and "great" prominently emerge. These terms strongly reflect customer satisfaction and endorsement, effectively capturing the essence of positive sentiment.

In contrast, the negative reviews reveal words such as "waste," "money," "don't," "buy," and "unfortunately," which accurately convey dissatisfaction and negative experiences. These terms align well with the expectations for negative sentiment, highlighting frustrations or regrets expressed by customers. The removal of noise and redundant information in the preprocessing phase enabled these sentiment-specific words to stand out more prominently.

The cleaned word clouds validate the effectiveness of the preprocessing steps, as sentiment-specific terms prominently emerge, accurately reflecting customer satisfaction or dissatisfaction. This improved clarity enhances the dataset's suitability for deeper analysis and modeling, ensuring sentiment patterns will be well-modeled.

4.2 Vectorization

Since computers process numerical data more effectively than textual data, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization method was employed to convert Amazon review text into numerical features for sentiment analysis. TF-IDF measures the importance of a word t in a specific review d, relative to its frequency across the entire corpus D of reviews. The TF-IDF score is calculated using the formula:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D),$$

where:

$$\text{TF}(t, d) = \frac{\text{Number of times } t \text{ appears in review } d}{\text{Total number of words in review } d},$$

$$\text{IDF}(t, D) = \log\left(\frac{|D|}{1 + |\{d \in D : t \in d\}|}\right).$$

Here, $|D|$ represents the total number of reviews (documents) in the dataframe, and $|\{d \in D : t \in d\}|$ is the number of reviews containing the word t. The addition of 1 in the denominator prevents division by zero for words that do not appear in any review.

By applying TF-IDF, the textual data is transformed into a numerical representation that highlights sentiment-specific patterns while reducing noise from frequent and less meaningful words. For example, after applying TF-IDF on a dataframe, terms like "excellent," "recommend," and "highly" are weighted heavily in positive reviews, whereas "poor," "return," and "waste" receive higher weights in negative reviews. This transformation enables the sentiment classification model to focus on meaningful patterns within the data.

The TF-IDF approach offers several key strengths when applied to Amazon reviews:

1. **Emphasizes Informative Words (Prediction)**: By assigning higher weights to rare but significant words, TF-IDF effectively captures terms like "refund," "waste," or "recommend," which are critical in identifying sentiment. This ensures that the sentiment analysis is driven by meaningful content rather than frequent, generic terms like "the" or "and."
2. **Reduces Noise from Common Words (Stability)**: Words like "product," "buy," and "use" often appear in Amazon reviews but do not inherently convey sentiment. TF-IDF assigns these words low weights, ensuring they do not dominate the analysis, thereby maintaining robust and interpretable predictions.
3. **Adaptability Across Product Categories (Stability)**: Amazon reviews cover diverse product categories, each with its unique vocabulary. TF-IDF adapts to these differences by dynamically weighting words based on their frequency in each dataset, making it suitable for a wide range of product reviews. This flexibility enhances its stability across diverse datasets.
4. **Interpretable Features (Prediction)**: The numerical features generated by TF-IDF are interpretable, allowing insights into the specific words that drive sentiment predictions. For example, a high weight for "durable" in a positive review of a product like a backpack provides clear evidence of customer satisfaction. This interpretability enhances understanding and predictability.
5. **Efficient Representation for Large & Sparse Data Representation (Computation)**: Given the Amazon review datasets often contain millions of entries, TF-IDF is computationally efficient and scales well, enabling quick transformation of large textual datasets into numerical form. The sparse matrix output of TF-IDF is also compact and memory-efficient, making it particularly well-suited for large datasets with many unique words across reviews.
6. **Scalability to High-Dimensional Data (Computation)**: TF-IDF efficiently handles the 1.9 million unique terms generated in this dataset, offering a balance between computational feasibility and meaningful representation of textual data. This scalability supports robust analysis even in high-dimensional feature spaces.

This approach was particularly important for this project due to the large number of features generated—approximately 1.9 million unique terms—capturing the breadth of vocabulary in the Amazon reviews. Managing such a high-dimensional feature space required a computationally efficient method like TF-IDF that represents the data while prioritizing sentiment-relevant terms, aligning well with the PCS framework to ensure predictive accuracy, computational efficiency, and stability across datasets.

5 Models

5.1 Pre-tuning Models

A diverse range of classification models was implemented and evaluated on the Amazon review sentiment dataset, guided by the Prediction, Computation, and Stability (PCS) framework. This framework emphasizes the balance between predictive performance (Prediction), computational feasibility (Computation), and robustness to perturbations in data and methodology (Stability). Each model was selected to address specific strengths and provide complementary insights under the PCS framework.

The modeling process began with a **Logistic Regression** model, chosen as a baseline due to its simplicity, interpretability, and computational efficiency. Logistic Regression aligns with the PCS framework as it offers high **computational efficiency** by scaling well to high-dimensional, sparse datasets like those generated by the TF-IDF vectorizer. Its assumption of a linear relationship between features and the target variable supports stable and predictable performance across different dataset partitions. Despite its simplicity, it provided a reliable benchmark for **prediction accuracy** and **stability**.

Next, the **Multinomial Naïve Bayes** model was employed, which is particularly effective for text classification tasks. This probabilistic model excels in **computation**, being lightweight and fast, making it a valuable tool for initial exploration of large datasets. However, its reliance on the independence assumption limited its **stability** in handling complex feature interactions, resulting in lower accuracy. To address non-linear relationships, a **Random Forest Classifier** was included, offering greater flexibility in capturing intricate patterns. Random Forest enhances **stability** by being robust to noise and overfitting, while its ability to measure feature importance supports interpretable and stable **prediction** outcomes.

Advanced models were incorporated to further optimize performance under the PCS framework. **XGBoost**, known for its powerful boosting techniques, iteratively refines **prediction accuracy** while maintaining robustness to noisy data, enhancing **stability**. Similarly, **LightGBM**, a gradient boosting algorithm, was included for its speed and scalability, addressing the **computation** aspect of PCS when handling large, sparse datasets like Amazon reviews. To explore neural network-based approaches, a **Multilayer Perceptron (MLP)** was implemented, balancing computational efficiency and flexibility. Finally, a **Long-Short Term Memory (LSTM)** model was employed, leveraging sequential patterns in textual data. LSTM's ability to capture long-range dependencies contributed to its strong **prediction** performance, but its high computational demands posed challenges for practical implementation, particularly in resource-constrained environments.

A.2 Table 1 summarizes the pre-tuning model performance. The **Logistic Regression** model emerged as the best-performing traditional approach, achieving a test accuracy of 90.15% and an F1 score of 93.09%. Its strong performance highlights the effectiveness of its linear assumptions in processing sparse TF-IDF features, while its **computational efficiency** ensured quick training times. This balance of **prediction** and **computation** made it a practical and robust choice for the dataset. Among traditional methods, **XGBoost** and **LightGBM** delivered competitive results, with XGBoost achieving the highest test accuracy (90.36%) and F1 score (93.14%) among ensemble methods, although at a higher computational cost. **Random Forest**, while robust and stable, was outperformed by these boosting methods, limiting its scalability for further optimization.

Neural network-based models demonstrated superior **prediction** performance, with the **LSTM** achieving the highest test accuracy (92.78%) and F1 score (94.45%). The LSTM model excelled in capturing sequential dependencies, enhancing its ability to process contextual information in reviews. However, its high **computational** demands and extended training times presented challenges for cost-sensitive or large-scale applications. The **MLP**, although slightly less accurate than LSTM, offered a more computationally efficient neural network alternative, achieving a test accuracy of 91.70%.

The **Multinomial Naïve Bayes** model struggled to achieve competitive **prediction accuracy**, with a test accuracy of 74.32% and an F1 score of 83.17%. Its limitations in handling complex feature interactions highlight the trade-offs in stability and accuracy inherent in its independence assumption.

Overall, the PCS framework guided the model evaluation process, balancing the strengths and trade-offs across prediction, computation, and stability. Logistic Regression emerged as the most practical choice, offering an optimal combination of **prediction performance** (93.09% F1 score),

computational efficiency, and stable outcomes. While advanced models like LSTM delivered higher accuracy, their resource demands reduced practicality in this case. Ensemble methods such as XGBoost and LightGBM provided robust alternatives but required more computational resources, but also leading to long training times.

5.2 Hyperparameter Tuning

Hyperparameter tuning was conducted on a subset of models, including Logistic Regression, Multinomial Naïve Bayes, XGBoost, and LightGBM, while excluding deep learning models and Random Forest due to their high computational demands and comparable performances. This decision was guided by the PCS framework, which ensures a balanced focus on predictive accuracy, computational feasibility, and robustness against variations in the data. The tuning process aimed to refine the models' performance while maintaining a manageable computational cost.

To optimize efficiency, random grid search was applied to 30% of the training data with a constrained parameter search space. This strategy balanced computational feasibility with sufficient data for meaningful optimization. After identifying the optimal parameters, the models were retrained on the full dataset and evaluated on the test data. Post-tuning results, summarized in A.2 Table 2, reveal the advantages and limitations of each model type and the impact of tuning.

Logistic Regression demonstrated marginal improvement post-tuning, achieving a test accuracy of 90.57% and an F1 score of 93.35%. This minimal change reflects its near-optimal default parameters and highlights its robustness as a baseline model. Its key strength lies in its computational efficiency and interpretability, which make it particularly well-suited for high-dimensional, sparse datasets like those generated by TF-IDF vectorization. However, its linear assumption limits its ability to capture complex, non-linear patterns in the data, which may constrain its performance in more intricate scenarios.

XGBoost, known for its powerful boosting techniques, maintained strong performance with a test accuracy of 90.36% and an F1 score of 93.16%. Its ability to handle high-dimensional data and capture non-linear relationships makes it a versatile choice for sentiment analysis. However, these benefits come at the cost of increased computational demands, particularly during the tuning phase, where multiple iterations and tree-building steps add to processing time. Despite these trade-offs, XGBoost's consistency across datasets underscores its stability and predictive reliability.

LightGBM exhibited a slight decline in performance after tuning, with a test accuracy of 75.88% and an F1 score of 84.21%. This reduction in predictive performance may stem from the reduced dataset size used during tuning and the constrained parameter search space, which limited its ability to fully optimize. While LightGBM is celebrated for its speed and efficiency, these results suggest that it may be sensitive to tuning conditions, impacting its stability in certain scenarios.

Multinomial Naïve Bayes experienced a further decline in performance, achieving a test accuracy of 72.46% and an F1 score of 81.90%. This reinforces its well-known limitation: the assumption of feature independence, which struggles with the complexities of sentiment analysis where interactions between words are critical. While computationally lightweight and fast, Multinomial Naïve Bayes is less capable of adapting to nuanced patterns in textual data, making it less effective for high-dimensional datasets.

From a PCS perspective, hyperparameter tuning provided limited benefits, with minimal gains in predictive accuracy and notable trade-offs in stability for some models. Logistic Regression stood out as the most practical and efficient model due to its robust baseline performance and minimal computational overhead, aligning well with the **computation** and **stability** aspects of the PCS framework. XGBoost emerged as a strong alternative, excelling in capturing complex patterns but requiring higher computational resources. LightGBM's sensitivity to tuning conditions highlighted potential instability, while Multinomial Naïve Bayes reaffirmed its limitations in handling complex datasets.

5.3 Areas of Improvement

Several strategies can be pursued to enhance model performance and address current limitations to refine feature representation, and explore more sophisticated models.

One significant area of improvement is **reducing feature sparsity**. The nearly 1.9 million TF-IDF features, while capturing a broad range of vocabulary, result in a high-dimensional feature space that can be computationally expensive and potentially redundant. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) or Truncated Singular Value Decomposition (SVD), could reduce sparsity while preserving essential information.

Another potential enhancement involves **incorporating n-grams** into the feature set. Bi-grams and tri-grams, such as "waste money" or "don't waste," can capture nuanced sentiment patterns that single-word features often miss. This would improve the model's ability to detect context-specific sentiment, particularly in phrases where word order is critical.

Incorporating **pretrained language models** represents another promising avenue for improvement. Models like BERT (Bidirectional Encoder Representations from Transformers) can capture contextual word meanings and relationships, addressing the limitations of simpler feature extraction methods. For instance, BERT can distinguish between polysemous words like "bank" (riverbank vs. financial institution) based on context, significantly improving the accuracy of sentiment classification. While computationally intensive, pretrained transformers have demonstrated state-of-the-art performance in text classification tasks and could be particularly valuable in refining sentiment analysis for Amazon reviews, hopefully boosting predictability.

6 Actionable Insights

A critical focus of this analysis was identifying the most influential words associated with negative sentiment in the Amazon product reviews. Extensive literature and market research underscore the importance of customer satisfaction as a key determinant of brand loyalty, both in monetary terms and psychological impact. Notably, a single negative review can often outweigh multiple positive ones in shaping customer perceptions and influencing purchasing decisions. To address this, the dataset was analyzed to uncover the most negatively polarized words, providing actionable insights into recurring themes of dissatisfaction.

The analysis began by identifying words with the most negative coefficients in the Logistic Regression model. These terms, visualized in a word cloud (Figure 4), included "poorly," "worst," "waste," "useless," and "disappoint." These words emerged as consistent markers of dissatisfaction, reflecting critical pain points for customers. This visualization provides a foundational understanding of recurring negative themes and their prevalence within the dataset.

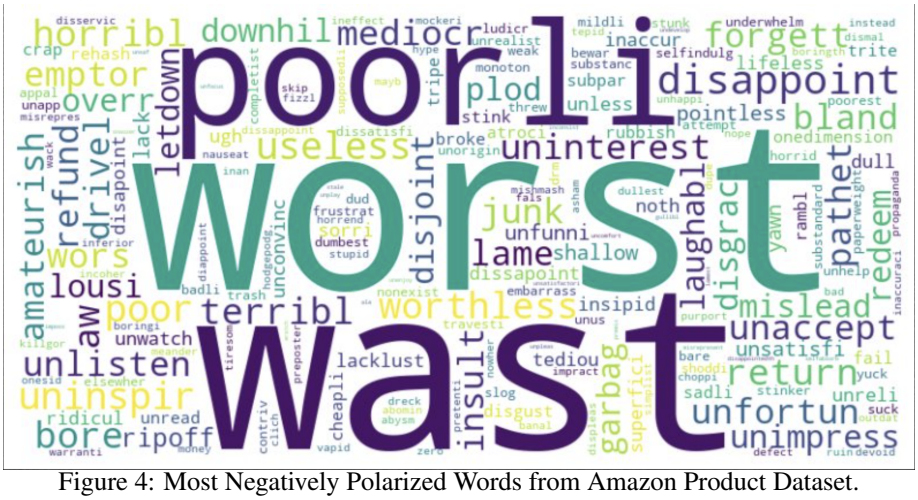


Figure 4: Most Negatively Polarized Words from Amazon Product Dataset.

To deepen this insight, a table (A.3 Figure 5) was created to contextualize these negatively polarized words further. The table provides detailed examples of the specific contexts in which these words appeared, offering valuable insights for both customers and sellers. Key unacceptable aspects, identified by customers, were emphasized (**bolded**) to draw attention to recurring issues. For instance, one entry revealed consistent complaints about Wi-Fi connectivity, underscoring a clear source of

frustration for customers. Another significant finding highlighted the impact of external factors on brand perception, such as the statement "Best Buy refused to help." This observation extends beyond product-specific issues, suggesting opportunities to collaborate with external partners to improve customer experiences and resolve logistical challenges.

This contextual analysis serves a dual purpose. For customers, it enhances transparency by highlighting common issues associated with certain products, empowering them to make informed purchasing decisions. For sellers, it provides a practical framework for addressing and resolving these pain points. By focusing on recurring themes of dissatisfaction, businesses can proactively enhance customer satisfaction, improve product quality, and refine after-sales support. Additionally, this analysis underscores the value of sentiment analysis as a tool for deriving actionable insights that inform not only product development but also broader customer service strategies.

These findings demonstrate the practical applications of sentiment analysis in identifying critical pain points, informing strategic decisions, and fostering collaboration between sellers and partners. By addressing these issues, businesses can strengthen brand loyalty, reduce negative reviews, and ultimately enhance the overall customer experience.

7 Future Enhancements/Work

Building upon the current analysis, several strategies can be implemented to enhance sentiment analysis and derive deeper insights from the dataset. These strategies focus on incorporating advanced techniques, broadening the scope of analysis, and enabling long-term monitoring of customer sentiment.

Topic Modeling for Thematic Analysis: Leveraging topic modeling techniques, such as Latent Dirichlet Allocation (LDA), can automatically identify and group the underlying themes or "topics" present in the text data. By scanning reviews and uncovering patterns of word usage that tend to cluster together, topic modeling enables the detection of recurring themes within negative sentiment reviews. This approach allows prioritization of work by focusing on larger, more prevalent topics in negative reviews, enabling targeted efforts to address critical pain points and improve customer satisfaction.

Product Category-Specific Analysis: Reintegrating omitted metadata from the test dataset, such as product categories (e.g., Kindle, Amazon Basics), can tailor the analysis to focus on nuanced feedback relevant to specific teams or departments. This refinement would provide insights that are granular and actionable for individual product lines, helping teams understand their unique challenges and opportunities. For example, issues flagged in the "Kindle" category could inform design improvements or feature enhancements, while feedback in the "Amazon Basics" category might reveal opportunities for improving quality control or packaging.

Tracking Long-Term Sentiment Trends: Developing models to monitor sentiment trends over time can provide valuable insights into recurring issues or emerging positive features across product categories. By analyzing temporal sentiment data, businesses can detect patterns such as a rise in complaints about a specific feature or growing customer appreciation for a newly introduced improvement. This long-term tracking would enable proactive responses to potential issues before they escalate and foster a deeper understanding of customer preferences.

8 Conclusion

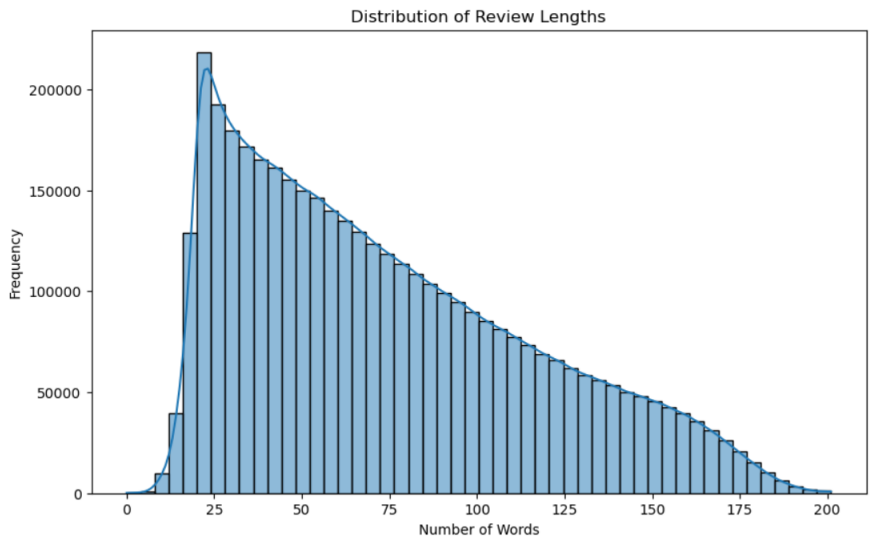
This analysis demonstrates the potential of sentiment analysis to go beyond surface-level metrics, such as star ratings, by uncovering nuanced insights from customer feedback. By identifying product-specific issues from detailed reviews, businesses can implement targeted improvements that directly address recurring pain points, thereby enhancing customer satisfaction. Additionally, sentiment analysis reveals minor but actionable complaints that may be overlooked in aggregated ratings, enabling companies to refine their offerings and address subtle concerns. Prioritizing product development efforts based on frequently praised or criticized features further ensures that resources are allocated effectively, focusing on areas that matter most to customers. These insights provide a roadmap for creating customer-centric solutions that not only improve product quality but also foster stronger brand loyalty and long-term competitive advantages.

9 References

1. https://github.com/alexbean55/AmazonReviews.
2. Training Data: https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews/data
3. Testing Data: https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products/datas
4. Amazon Web Services (AWS). *What is Sentiment Analysis?*. Available at: https://aws.amazon.com/what-is/sentiment-analysis/.
5. Repustate Blog. *Sentiment Analysis Real-World Examples*. Available at: https://www.repustate.com/blog/sentiment-analysis-real-world-examples/#:~:text=and%20competitor%20analysis.-,What%20is%20Sentiment%20Analysis?,Learn%20About%20Sentiment%20Analysis%20Applications./.
6. Towards Data Science. *Multi-Class Text Classification with Scikit-Learn*. Available at: https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f.
7. Li, Susan. *Fake News Classification with LSTM or Logistic Regression*. Available at: https://actsusanli.medium.com/fake-news-classification-with-lstm-or-logistic-regression-82a3527aaf13.
8. Towards Data Science. *Multi-Class Text Classification with LSTM*. Available at: https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17.
9. Analytics Vidhya. *How to Generate Word Clouds from Multiple News Sources with Python Code*. Available at: https://medium.com/analytics-vidhya/how-to-generate-word-clouds-from-multiple-news-source-with-python-code-47a03c512fe4.

A Appendix / Supplemental Material

A.1 EDA: All Data Histogram



A.2 Modeling: Pre-tuning model performance

Model	Test Accuracy	Test F1 Score
Logistic Regression	0.9015	0.9309
Multinomial Naïve Bayes	0.7432	0.8317
Random Forest Classifier	0.8161	0.8786
XGBoost Classifier	0.9036	0.9314
LightGBM Classifier	0.8979	0.9286
MLP	0.9170	0.9245
Long-Short Term Memory	0.9278	0.9445

Table 1: Pre-tuning model performance on the Amazon review test dataset.

A.3 Modeling: Pre-tuning model performance

Model	Test Accuracy	Test F1 Score
Logistic Regression	0.9057	0.9335
Multinomial Naïve Bayes	0.7246	0.8190
XGBoost Classifier	0.9036	0.9316
LightGBM Classifier	0.7588	0.8421

Table 2: Post-tuning model performance on the Amazon review test dataset.

A.4 Actionable Insights: Most Negative Reviews

Waste of money
Returned it. it was a **poor reader** and clearly **would not last**..
Could not get this to **work with my chromecast** so I returned it.
I returned the item was not what I expected very disappointed white the table
Not sure why, but it kept **freezing** and ended up having to return it. Very disappointing
This product not good waste of money I would recommend it
Worst product ever. Constantly screwing up and **filled with ads** you have to pay to remove. **Best Buy refused to help** when I wanted to return it
Didn't work more that 4 months. Don't waste your money.
Truly I should have tried to get a refund. I threw out the packaging before I realized how poorly it works. This **charger** only works every third day when held at a particular angle. The planets almost need to align for it to charge my Kindle. Complete disappointment.
Absolute junk, don't waste your money. Bought for wife for a Christmas gift. **Would not connect to Wi-Fi** at my house. Useless!

Figure 5: Top 10 Most Negative Reviews for Amazon Products