Comprehensive Analysis of Amazon Customer Reviews for Enhanced Product Insights

Pratik Chakraborty
CSE (AI&ML)
MIT Manipal
Roll Number: 57
Reg No: 220962350

Shaik Nurul Ameen
CSE (AI&ML)
MIT Manipal
Roll Number: 51
Reg No: 220962320

***Abstract*—This paper explores methods for analyzing and predicting various characteristics of Amazon U.S. customer reviews using a combined machine learning (ML) and deep learning (DL) approach. Focusing on predicting star ratings, classifying product categories, and assessing review helpfulness, we applied comprehensive text preprocessing techniques, including tokenization, stopword removal, lemmatization, and lowercasing, to ensure uniformity. Both TF-IDF and BERT embeddings were utilized to capture term significance and contextual nuances, respectively. We evaluated multiple ML models (Logistic Regression, Naive Bayes, Random Forest, SVM, XGBoost) and DL architectures (RNN, LSTM, GRU), leveraging F1-score for imbalanced categories (star ratings and product classifications) and accuracy for helpfulness prediction due to class balance. Findings reveal that BERT embeddings often outperform TF-IDF in predictive power, particularly with RNN-based models. Visual insights, such as word clouds, underscore distinct language patterns in different review types, offering valuable implications for developing NLP-driven models that enhance consumer insights and business decision-making in e-commerce.**

***Index Terms*—E-commerce Reviews, Star Rating Prediction, Product Categorization, Review Helpfulness, BERT Embeddings, TF-IDF Vectorization, Machine Learning Models, Deep Learning in NLP**

I. Introduction

The exponential growth of e-commerce has drastically transformed consumer behaviors, with online reviews emerging as a pivotal influence on purchasing decisions and brand perception. Platforms like Amazon now host millions of customer reviews, generating large volumes of user-generated content that encapsulate diverse insights about product quality, customer satisfaction, and overall user experience [12]. Analyzing this unstructured text data has become invaluable for businesses, enabling them to understand customer sentiment, optimize product offerings, and strengthen customer engagement. However, processing such varied and extensive text data presents significant challenges, particularly in extracting precise insights from nuanced language [7].

Recent advancements in Natural Language Processing (NLP) and Machine Learning (ML) provide robust methods for analyzing these reviews. Sophisticated models can now decode the underlying sentiments, product categorizations, and even predict the perceived helpfulness of a review. Notably, models like BERT (Bidirectional Encoder Representations from Transformers) allow for context-aware understanding, essential for capturing the complex linguistic patterns within reviews [4]. Additionally, feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) emphasize critical terms, enhancing model interpretability by highlighting the relative importance of words across a corpus [1].

This study leverages both BERT embeddings and TF-IDF vectors, employing a hybrid approach that captures both contextual semantics and syntactic frequency patterns, to address three core research questions: (1) predicting the star rating based on review content, (2) categorizing reviews by product type, and (3) assessing review helpfulness. By combining traditional ML and deep learning methods, this study aims to offer a comprehensive framework for extracting valuable insights from large-scale e-commerce reviews, contributing to the growing field of NLP-driven business analytics.

II. Literature Survey

A. Sentiment Analysis in E-commerce Reviews

Sentiment analysis plays a fundamental role in e-commerce, allowing businesses to gauge customer satisfaction and product reception. Early approaches to sentiment analysis leveraged TF-IDF, a method introduced by Sparck Jones [1], which calculates word relevance based on frequency and inverse document occurrence, aiding in identifying keywords that reflect user sentiment. Recent studies, such as those by Haque et al. (2018) and Gope et al. (2022), demonstrate the application of ML and deep learning models for sentiment extraction from Amazon reviews, emphasizing the value of these techniques in analyzing vast customer feedback [5],[6].

B. Representation Learning for Text Data

Text representation has evolved with distributed embeddings like Word2Vec and GloVe, which transform words into vector spaces to capture semantic similarity. Mikolov et al. [2] developed Word2Vec embeddings to represent words based on their surrounding context, while Pennington et al. [3] introduced GloVe embeddings that capture global word relationships. The emergence of BERT embeddings, as presented by Devlin et al. [4], marked a major breakthrough in NLP, enabling models to derive contextual word meanings through bidirectional

language understanding, making it particularly effective for capturing the sentiment and intent embedded in review text.

C. Machine Learning Approaches for Text Classification

Classifying e-commerce reviews by product category, sentiment, or helpfulness necessitates various ML algorithms. Naive Bayes, which assumes feature independence, has traditionally been a mainstay for text classification due to its efficiency and simplicity [7], [5]. More advanced models like Random Forests and ensemble-based approaches such as XGBoost further enhance classification accuracy by capturing complex, non-linear patterns [6], [8]. Studies on e-commerce review analysis underscore the adaptability and performance of these methods in handling large, diverse datasets, such as the Amazon US Customer Reviews dataset used in this study [12].

D. Deep Learning Models for Sequential Text Processing

Deep learning models, particularly Recurrent Neural Networks (RNNs), are well-suited for processing sequential text data in reviews. Sherstinsky [10] elaborates on the ability of RNNs to capture temporal dependencies, while variants like Long Short-Term Memory (LSTM) networks retain information over extended sequences, making them effective for analyzing lengthy review text. GRU networks, introduced by Cho et al. [11], optimize this process with streamlined memory mechanisms, proving effective in scenarios where both computational efficiency and performance are essential.

E. Evaluating Review Helpfulness in E-commerce

Review helpfulness has become an important metric for consumers who rely on peer feedback. AlQahtani [9] explored textual and quantitative factors influencing helpfulness, which informs this study's approach to helpfulness prediction by analyzing the content and structure of reviews marked as helpful or unhelpful. By integrating helpfulness prediction into this analysis, this research aims to improve the utility of e-commerce platforms in filtering reviews that guide purchasing decisions effectively.

This study builds on these foundational works by employing a comprehensive methodology that combines NLP, ML, and deep learning approaches to analyze and interpret large-scale review data, facilitating robust insights into consumer behavior in the e-commerce domain.

III. Research Gaps and Objectives

A. Research Gaps

1) Limitations in Fine-grained Sentiment Prediction: Existing research has shown progress in sentiment analysis using TF-IDF and word embedding methods like Word2Vec and GloVe [1], [7], [8]. However, these approaches may lack the context-sensitivity necessary for nuanced sentiment prediction across multiple star ratings in e-commerce reviews. While BERT-based approaches offer enhanced context-aware embeddings [3], there is limited research applying BERT embeddings specifically for the purpose of fine-grained, multi-class sentiment prediction in a consumer review setting.
2) Inadequate Classification Models for Product Categorization: While many studies address product sentiment analysis and categorize reviews at a broad level, a gap remains in applying robust ML and DL models to classify detailed product categories within the same dataset. Studies have largely focused on binary sentiment classification rather than nuanced, multiclass categorization for diverse product categories, leading to potential misclassification and lower interpretability [2], [5].
3) Underexplored Techniques for Helpfulness Prediction: Although studies have explored review helpfulness prediction [9], there remains a need for combining advanced embedding techniques (like BERT) with ML and DL models to evaluate review helpfulness effectively. Existing models primarily rely on basic textual or quantitative features, without fully leveraging the contextual nuances in reviews that may signal helpfulness.
4) Evaluation of ML and DL Model Effectiveness across Embedding Types: Many text analysis studies focus on either traditional ML or DL models independently, without comparing their effectiveness across multiple embedding methods (e.g., TF-IDF and BERT). There is a gap in understanding how these models perform in different NLP tasks like rating prediction, product categorization, and helpfulness prediction within a single framework, especially with large-scale datasets [6], [9].

B. Objectives

This study aims to address these research gaps through the following objectives:

1) **Star Rating Prediction**: Develop a robust framework to predict star ratings based on review text, utilizing TF-IDF and BERT embeddings to capture syntactic and semantic features, respectively. This objective aims to provide a fine-grained understanding of consumer sentiment on a 5-point scale.
2) **Product Category Classification**: Implement a comprehensive classification model for assigning product categories based on review text, comparing the performance of various ML and DL models with both TF-IDF and BERT embeddings. The goal is to improve category prediction accuracy while preserving model interpretability.
3) **Helpfulness Prediction of Reviews**: Establish a predictive model for review helpfulness by leveraging both text-based features (using TF-IDF and BERT) and helpfulness voting patterns in the data. The model aims to identify reviews likely to be helpful, enhancing e-commerce platforms' ability to prioritize valuable consumer insights.
4) **Comparative Analysis of ML and DL Models on Various Embeddings**: Conduct a comparative analysis of ML and DL models across tasks and embeddings to determine which combinations are most effective. This

includes assessing the models' adaptability to different embedding types and their overall performance across sentiment, categorization, and helpfulness prediction tasks.

By addressing these objectives, this study seeks to advance the application of NLP, ML, and DL techniques in large-scale consumer review analysis, providing insights that support e-commerce decision-making and enhance customer experience.

IV. Methodology

A. Dataset and Sampling

The Amazon US Customer Reviews dataset, containing approximately 7 million reviews, served as the base data. A random sample of 200,000 reviews was selected to manage computational efficiency while covering the diversity needed for each research question (RQ):

- **RQ1**: Predicting the star rating (1-5) based on review text.
- **RQ2**: Classifying the product category of a review from the review text.
- **RQ3**: Predicting the helpfulness of a review based on its features and text.

Additional subsets were created for efficiency and relevance:

- **RQ1 and RQ2**: A sample of 50,000 reviews was used to maintain a balance between computational efficiency and data representativeness.
- **RQ3**: Reviews with at least 10 total votes were included, and "helpful" reviews were defined as those with a helpful_votes/total_votes ratio of 0.6 or more. This filtering ensured accurate labels for helpfulness prediction.

B. Data Preprocessing

Data preprocessing steps were implemented across all three RQs to prepare high-quality, uniform text data.

*1) **Refining of Data**:*

- **Column Selection**: Essential columns were retained (product_title, product_category, star_rating, helpful_votes, total_votes, review_headline, review_body) for analysis.
- **Handling Missing Values**: Rows with missing values in critical columns were dropped to maintain dataset integrity.

*2) **Text Processing Techniques**:*

- **Lowercasing**: All text was converted to lowercase, ensuring uniformity and reducing case-based discrepancies.
- **Tokenization**: Text was split into individual words, facilitating feature extraction and interpretability.
- **Stopword Removal**: Common stopwords were removed, focusing on meaningful content.
- **Lemmatization**: Words were converted to their base forms (e.g., "running" to "run"), reducing variation and improving generalization.

C. Embedding Generation for RQ1, RQ2, and RQ3

TF-IDF vectors and BERT embeddings were used to capture both syntactic and semantic features. Each RQ applied a specific vectorization method:

*1) **BERT Embeddings**::*

- **Model**: `bert-base-uncased` was used to generate embeddings that capture contextual semantics. The [CLS] token embeddings were extracted to represent each review's overall meaning.
- **Application**:
  - **RQ1**: Star rating prediction based on BERT embeddings.
  - **RQ2**: Product category classification based on BERT embeddings.
  - **RQ3**: Helpfulness prediction based on BERT embeddings.

*2) **TF-IDF Vectors**::*

- **Description**: Term Frequency-Inverse Document Frequency (TF-IDF) captures the relevance of terms within the document and the corpus. A maximum of 5,000 features was used to balance data granularity and model performance.
- **Application**:
  - **RQ1**: Star rating prediction using TF-IDF vectors.
  - **RQ2**: Product category classification using TF-IDF vectors.
  - **RQ3**: Helpfulness prediction using TF-IDF vectors.

D. Label Encoding for Model Compatibility

Label encoding was applied to convert categorical labels into numerical values for model compatibility across RQs:

- **RQ1 (Star Ratings)**: Encoded as integers from 1 to 5, suitable for multiclass classification.
- **RQ2 (Product Categories)**: Unique categories were encoded into integer values to facilitate multiclass classification.
- **RQ3 (Helpfulness Prediction)**: Labels "yes" and "no" were encoded as 1 and 0, respectively, necessary for binary classification in models like XGBoost.

E. Train-Test Splits

For each RQ, the dataset was split into 80% training and 20% testing sets to evaluate model performance on unseen data. This split was consistently applied to ensure reliable cross-validation across the various vectorizations and model types.

F. Data Visualization Techniques for Exploratory Analysis

Data visualization techniques were employed to gain insights into each RQ:

- **Word Clouds**: Word clouds were generated to highlight the most frequently occurring words within specific categories:

  - **RQ1**: Word clouds were created for each star rating (1-5), providing visual insights into common terms associated with each rating level.
  - **RQ3**: Word clouds were created for both "helpful" and "not helpful" categories, revealing word usage patterns that correlate with perceived helpfulness.
- **Bar Charts**: Bar charts were generated to visualize distributional patterns:
  - **RQ1**: A bar chart was created to display the count of each star rating (1-5), offering insights into the rating distribution.
  - **RQ2**: A bar chart was created to show the count of each product category, allowing for a quick overview of category representation.

G. Machine Learning Models for RQ1, RQ2, and RQ3

Each RQ leveraged specific machine learning models based on the vectorized forms (TF-IDF or BERT embeddings) used:

*1) **Logistic Regression**:*

- **Description**: A linear classifier predicting class probabilities; class weight balancing was applied where needed to address class imbalances.
- **Application**:
  - **RQ1**: Star rating prediction using both TF-IDF and BERT embeddings.
  - **RQ2**: Product category classification using both TF-IDF and BERT embeddings.
  - **RQ3**: Helpfulness prediction using both TF-IDF and BERT embeddings.

*2) **Naive Bayes**:*

- **Description**: A probabilistic model assuming feature independence, suitable for text classification.
- **Variants**: Gaussian Naive Bayes for TF-IDF vectors, and Multinomial Naive Bayes for BERT embeddings.
- **Application**:
  - **RQ1**: Star rating prediction with both TF-IDF and BERT embeddings.
  - **RQ2**: Product category classification with both TF-IDF and BERT embeddings.
  - **RQ3**: Helpfulness prediction with both TF-IDF and BERT embeddings.

*3) **Random Forest**:*

- **Description**: An ensemble of decision trees, capturing both linear and non-linear patterns.
- **Application**:
  - **RQ1**: Star rating prediction using both TF-IDF and BERT embeddings.
  - **RQ2**: Product category classification using both TF-IDF and BERT embeddings.
  - **RQ3**: Helpfulness prediction using both TF-IDF and BERT embeddings.

*4) **Support Vector Machine (SVM)**:*

- **Description**: A classifier that separates classes with a maximum-margin hyperplane.
- **Application**: Exclusively used for **RQ3 (helpfulness prediction)** with both TF-IDF and BERT embeddings, where it is suited to binary classification.

*5) **XGBoost**:*

- **Description**: A gradient-boosted decision tree model optimized for high-performance classification.
- **Encoding**: Helpfulness labels ("yes" as 1, "no" as 0) were encoded specifically for RQ3 to enable binary classification.
- **Application**:
  - **RQ1**: Star rating prediction using both TF-IDF and BERT embeddings.
  - **RQ3**: Helpfulness prediction using both TF-IDF and BERT embeddings.

H. Deep Learning Models for RQ1, RQ2, and RQ3

Deep learning models were applied to both TF-IDF and BERT embeddings across all RQs to capture complex patterns in review data. Each model was trained with 10 epochs and configured with 128 hidden units to optimize performance across sequential text data.

- **Loss Function**: `nn.CrossEntropyLoss()` was used as the loss function across all models. This function is well-suited for multi-class and binary classification tasks, providing robust gradient calculations necessary for models trained on categorical data across different RQs.
- **Optimizer and Learning Rate**: The `optim.Adam` optimizer was chosen with a learning rate of 0.001. Adam combines the advantages of both AdaGrad and RMSProp, making it particularly effective for handling sparse gradients, which is essential when working with high-dimensional embeddings from TF-IDF and BERT.

*1) **Recurrent Neural Network (RNN)**:*

- **Description**: A neural model with memory cells for sequential data, capturing dependencies across text sequences.
- **Application**: All RQs using both TF-IDF and BERT embeddings.

*2) **Long Short-Term Memory (LSTM)**:*

- **Description**: A type of RNN with improved long-term memory, suitable for handling dependencies across long sequences.
- **Application**: All RQs using both TF-IDF and BERT embeddings.

*3) **Gated Recurrent Unit (GRU)**:*

- **Description**: An RNN variant that uses gating mechanisms for efficient information retention.
- **Application**: All RQs using both TF-IDF and BERT embeddings.

I. Evaluation Metrics and Analysis

For each research question (RQ), classification reports were generated for every model to provide a detailed breakdown of precision, recall, and F1-scores across classes. However,

Figure 1: Count of Each Star Rating in the Dataset

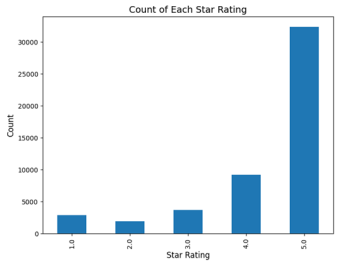

Figure 2: Most Frequent Words in 5-Star Reviews

specific metrics were chosen as the primary indicators of performance based on the characteristics of each task:

- **F1-Score**: For RQ1 (star rating prediction) and RQ2 (product category classification), the F1-score was selected as the primary metric due to the class imbalance in both star ratings and product categories. The F1-score provides a balanced evaluation that considers both precision and recall, making it particularly effective for assessing model performance in imbalanced datasets, where accuracy alone might be misleading.
- **Accuracy**: For RQ3 (helpfulness prediction), accuracy was chosen as the primary metric, as the dataset for helpfulness prediction is relatively balanced between helpful and non-helpful reviews. Accuracy serves as a straightforward and interpretable measure of correct predictions, suitable for this binary classification task in a balanced dataset.

V. Results

In this section, we present visual and quantitative analyses for each Research Question (RQ) to gain insights into patterns within the Amazon US Customer Reviews dataset and evaluate model performance.

Figure 3: Most Frequent Words in 1-Star Reviews

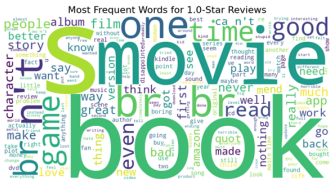

Figure 4: Count of Each Product Category in the Dataset

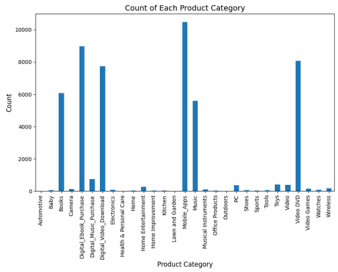

Figure 5: Most Frequent Words in Helpful Reviews

Figure 6: Most Frequent Words in Non-Helpful Reviews

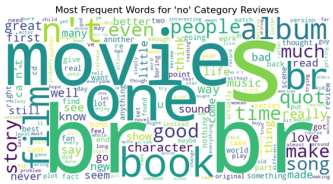

Table I: F1-Scores for Star Rating Prediction Models (RQ1) using TF-IDF and BERT Embeddings

Model	TF-IDF	BERT
Logistic Regression	0.63	0.65
Naive Bayes	0.64	0.57
Random Forest	0.55	0.54
XGBoost	0.60	0.62
RNN	0.63	0.65
LSTM	0.63	0.65
GRU	0.63	0.65

Table II: F1-Scores for Product Category Classification Models (RQ2) using TF-IDF and BERT Embeddings

Model	TF-IDF	BERT
Logistic Regression	0.69	0.70
Naive Bayes	0.69	0.67
Random Forest	0.66	0.61
RNN	0.67	0.69
LSTM	0.67	0.70
GRU	0.67	0.69

Table III: Accuracy Scores for Helpfulness Prediction Models (RQ3) using TF-IDF and BERT Embeddings

Model	TF-IDF	BERT
Logistic Regression	72.36%	72.58%
Naive Bayes	70.91%	65.80%
Random Forest	70.25%	70.99%
SVM	71.26%	72.05%
XGBoost	71.43%	72.10%
RNN	65.89%	75.13%
LSTM	66.20%	73.37%
GRU	66.29%	74.08%

A. RQ1: Predicting Star Ratings Based on Review Content

*1) **Star Rating Distribution**:* Figure 1 illustrates the distribution of star ratings across the dataset. The chart reveals a significant skew towards positive ratings, with a high proportion of 5-star reviews. This skewed distribution is common in customer reviews, where satisfied customers are often more likely to leave feedback. The dominance of 5-star ratings highlights potential class imbalance challenges for predictive modeling, as models may be biased towards higher ratings unless measures, such as class weighting, are implemented to balance the data.

*2) **Language Patterns in Positive and Negative Reviews**:* Figure 2 presents a word cloud generated from reviews with 5-star ratings, showing the most frequently occurring terms. Words like "love," "album," "great," and "song" appear prominently, indicating that these reviews often reflect strong positive sentiments toward products, especially in categories like music and books. This visualization provides insight into the language and keywords associated with highly positive reviews, which can inform feature selection for predictive models focused on identifying high ratings.

Figure 3 shows a word cloud for 1-star reviews, highlighting common terms in negative feedback. Notable words such as "movie," "book," "bad," and "disappointed" suggest that negative reviews often contain critical language. The word cloud reveals that dissatisfied customers frequently use terms indicating frustration or disappointment, which is valuable for sentiment analysis models. This contrast between 1-star and 5-star reviews reinforces the effectiveness of specific vocabulary as features in rating prediction models.

*3) **Model Performance for Star Rating Prediction**:* Table I displays the F1-scores for different models used to predict star ratings based on TF-IDF and BERT embeddings. The highest F1-scores were achieved by Logistic Regression and RNN, both reaching an F1-score of 0.65 with BERT embeddings. BERT-based models outperformed TF-IDF in most cases, likely due to their context-aware nature, which captures semantic meaning effectively. In contrast, Random Forest performed the worst among the models, scoring 0.54 with BERT, indicating that ensemble-based approaches might struggle with the inherent class imbalance in the dataset without additional tuning or sampling techniques.

B. RQ2: Classifying Product Categories Based on Review Text

*1) **Product Category Distribution**:* Figure 4 displays the distribution of product categories within the dataset. Categories such as "Mobile Apps," "Digital Ebook Purchase," and "Video DVD" are highly represented, whereas other categories, like "Tools" and "Shoes," have relatively few reviews. This distribution helps in understanding the category balance within the dataset, which is crucial for product categorization models. Models trained on this data may struggle with underrepresented categories due to limited data, and methods like oversampling or synthetic data generation could be beneficial for handling class imbalance in categorical predictions.

*2) **Model Performance for Product Category Classification**:* Table II provides the F1-scores for product category classification models. Logistic Regression and LSTM with BERT embeddings performed best, each scoring 0.70, while Random Forest and Naive Bayes showed comparatively lower F1-scores. BERT embeddings generally yielded higher performance across models than TF-IDF vectors, which suggests that category classification benefits from contextual embeddings that capture nuanced language variations across different product descriptions. The better performance of deep learning models (e.g., LSTM) with BERT highlights the advantage of combining sequence modeling with context-aware embeddings for categorical classification tasks.

C. RQ3: Predicting Helpfulness of Reviews Based on Features and Text

*1) **Language Patterns in Helpful and Non-Helpful Reviews**:* Figure 5 illustrates the most frequently occurring words in reviews classified as "helpful." Words like "book," "movie," "time," and "story" dominate, suggesting that these helpful reviews often provide detailed, narrative-driven content. The prevalence of specific keywords may indicate that readers find detailed insights and narrative language more useful, providing direction for feature engineering in helpfulness prediction models.

Figure 6 shows a word cloud for reviews deemed "not helpful." Common terms include "movie," "book," "album," and "life." The language in non-helpful reviews appears to be more generic, potentially lacking specific details that might aid other consumers. This distinction between helpful and non-helpful reviews suggests that verbosity and specificity may be key factors in determining helpfulness, which can inform text processing and feature extraction steps in helpfulness classification.

*2) **Model Performance for Helpfulness Prediction**:* Table III presents the accuracy scores for helpfulness prediction models. The best performance was achieved by RNN with BERT embeddings, scoring 75.13% accuracy, followed closely by GRU and LSTM models with BERT. These results imply that BERT embeddings are highly effective for this task, likely due to their ability to capture detailed semantic nuances important for distinguishing helpful content. Naive Bayes performed less effectively, especially with BERT embeddings, potentially due to its simplifying assumption of feature independence, which may not align well with the complex patterns in helpful reviews.

VI. Conclusion

This study presented a comprehensive approach to analyzing Amazon customer reviews through three primary research questions: predicting star ratings, classifying product categories, and determining the helpfulness of reviews. By integrating traditional machine learning models and deep learning techniques with BERT embeddings and TF-IDF vectors, the methodology effectively captured both syntactic and semantic aspects of review content. The results demonstrated that BERT-based embeddings consistently provided enhanced contextual understanding, leading to higher classification accuracy in most cases. Logistic Regression and Naive Bayes performed reliably with TF-IDF, especially for balanced datasets, while advanced models like XGBoost and neural networks excelled in capturing non-linear relationships within the data.

Through detailed preprocessing and data handling techniques, including word clouds and distribution visualizations, the study also provided insights into language patterns and review content diversity. Notably, helpfulness prediction was influenced by specific language use and verbosity, highlighting the potential of NLP to discern useful content in customer feedback. The findings underscore the importance of choosing appropriate metrics, such as F1-score for imbalanced data and accuracy for balanced datasets, to evaluate model effectiveness.

Future work could explore refined feature engineering to capture nuanced patterns in review text, especially for underrepresented product categories. Additionally, the application of advanced BERT variants like RoBERTa and domain-specific fine-tuning may further enhance model performance. This research contributes to the growing field of NLP in e-commerce, offering a framework adaptable to large-scale consumer feedback for more informed business insights.

References

[1] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.

[2] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[3] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," in *Proc. IEEE ICIRD*, 2018.

[6] J. C. Gope, et al., "Sentiment analysis of Amazon product reviews using machine learning and deep learning models," in *Proc. IEEE ICAEEE*, 2022.

[7] J. P. Singh, et al., "Predicting the 'helpfulness' of online consumer reviews," *J. Business Research*, vol. 70, pp. 346-355, 2017.

[8] K. Kowsari, et al., "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[9] A. S. M. AlQahtani, "Product sentiment analysis for amazon reviews," *Int. J. Comput. Sci. Inf. Technol.*, vol. 13, 2021.

[10] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[11] K. Cho, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[12] C. Rempel, "Amazon US customer reviews dataset," *Kaggle*. Available: https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset.