Sentiment Analysis and Topic Modeling of Amazon Product Reviews

Abdul Rehman (417997), Hamza Mahmood (413603), and Abdul Rehman (408651)

***Abstract*— Sentiment Analysis is the computational process of identifying and categorizing the tone and emotional context of statements within a text to determine whether the author's perspective on a specific product is positive or negative. This analysis is crucial for industry leaders like Amazon, whose vast e-commerce operations rely on understanding consumer behavior and sentiment to enhance customer engagement. With advancements in machine learning and text analytics, this task can now be accomplished effectively. In this project, we aim to predict the sentiment of reviews for Amazon Electronics products using supervised and unsupervised machine learning algorithms. We will also evaluate and compare the performance of the approaches employed to determine the most effective method for such applications.**

***Index Terms*—Classification, Feature extraction, Sentiment analysis, Word embedding.**

I. Introduction

The rapid growth of e-commerce has resulted in the generation of vast amounts of data, which presents both opportunities and challenges for online platforms. Understanding customer sentiments, especially through product reviews, plays a pivotal role in shaping business decisions, improving customer experience, and refining product offerings. This project leverages the Amazon Electronics Product Review Dataset, containing over 1.6 million reviews, to explore sentiment analysis and topic modeling. By employing both supervised learning techniques, such as sentiment classification based on ratings, and unsupervised learning techniques like Latent Dirichlet Allocation (LDA) for clustering reviews, the project aims to derive actionable insights from customer feedback. Various classification models, including CountVectorizer and TF-IDF, are implemented to evaluate the most effective sentiment prediction approach. The ultimate goal is to provide e-commerce platforms with a robust framework for analyzing and categorizing customer opinions, helping them make data-driven decisions to enhance product offerings and customer satisfaction.

n the digital age, online platforms like Amazon have become the go-to destination for millions of consumers worldwide, making them a primary source of product reviews and feedback. As a result, understanding the sentiment behind these reviews is crucial for companies aiming to stay competitive and improve their services. Sentiment analysis, the process of categorizing text into positive, negative, or neutral sentiments, has emerged as a powerful tool for analyzing customer opinions and behaviors. In particular, product reviews on e-commerce platforms offer valuable insights into consumer preferences, pain points, and satisfaction levels.

Amazon, as one of the largest e-commerce platforms globally, hosts an enormous amount of customer feedback on its products. This feedback is often expressed through reviews, which include both textual descriptions and ratings. However, manually interpreting these reviews is a labor-intensive and time-consuming task. This project addresses this challenge by automating the sentiment analysis of product reviews using machine learning techniques. The Amazon Electronics Product Review Dataset, consisting of over 1.6 million entries, is the focal point of this study. By applying both supervised learning techniques for sentiment classification and unsupervised methods for clustering reviews, this project aims to uncover meaningful patterns within the reviews that can be used to better understand customer sentiment.

The motivation behind this project is twofold. First, by leveraging sentiment analysis, we aim to create an automated framework that can efficiently process large volumes of customer reviews, thus offering businesses the ability to monitor and respond to customer feedback in real-time. Second, by incorporating topic modeling through LDA, we hope to identify emerging trends and key issues that customers frequently discuss, providing businesses with actionable insights into customer expectations and areas for improvement. Ultimately, this research aims to contribute to the growing field of customer sentiment analysis, offering both a technical solution and strategic insights for e-commerce platforms striving to enhance their customer relationships.

Figure 1 illustrates the workflow of our project. We begin by preprocessing the data, followed by the application of the Bag-of-Words approach, where both Count Vectorizer and TF-IDF methods are implemented. We also consider various classifiers for sentiment prediction. For clustering, we apply Latent Dirichlet Allocation (LDA) to identify dominant topics within the reviews. Finally, we compare and evaluate the performance of sentiment classification and clustering methods to determine the most effective approaches for understanding customer sentiment in this context.

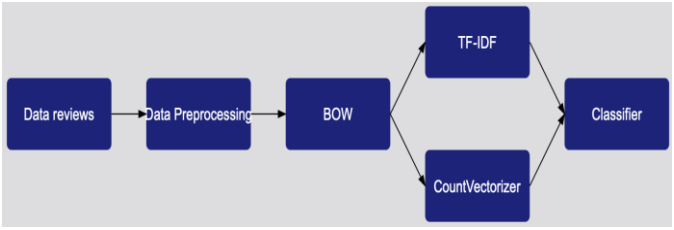


Fig. 1: Project Overflow.

II. Literature Review

As part of the literature review and previous research in the field of sentiment analysis, we examined the work of Pang et al. [8], which was one of the earliest attempts to classify documents based on sentiment rather than topics. They noted that traditional machine learning methods such as Naive Bayes, Maximum Entropy Classification, and Support Vector Machines (SVM) do not perform as well on sentiment classification compared to traditional topic-based categorization.

However, we found several studies that contradicted these claims. For instance, in [9], the author applied existing supervised learning algorithms to the Yelp dataset, including Naive Bayes, Perceptron, and Multiclass SVM, and compared their predictions with the actual ratings. The study concluded that binarized Naive Bayes, combined with feature selection, stop word removal, and stemming, was the most effective for sentiment analysis on such datasets. It is worth noting that the features of the Yelp dataset closely resemble those of the Amazon product review dataset used in this project.

Additionally, we found a paper [3] that addresses the fundamental challenge in sentiment analysis: sentiment polarity categorization. The study considers both review-level and sentence-level categorization. The authors preprocessed the data by extracting all subjective content, i.e., sentences containing at least one positive or negative word. Negative phrases were identified using POS tagging and negative prefixes, followed by sentiment score computation. For training the classifiers, each data entry was transformed into a feature vector with binary strings representing tokens of words in the sentence. They applied 10-fold cross-validation, using sentiment scores to identify positive and negative classes. The sentiment score proved to be a strong feature, achieving an F1 score of 0.73 for review-level categorization and 0.8 for sentence-level categorization. However, the study noted limitations, such as poor performance when F1 scores were very low and when dealing with implicit sentiments. The classification models used included Naive Bayes, Decision Trees, and Support Vector Machines.

In addition, [5] employed a supervised learning approach to label an unlabeled dataset using active learning. The data preprocessing included tokenization, stop word removal, and POS tagging. The authors used a combination of two approaches for feature extraction and evaluated the classification performance based on Precision, Recall, F-measure, and Accuracy. The study concluded that SVM performed the best when a large number of datasets were available.

By reviewing these works, we aim to validate their findings by applying some of their approaches to our dataset to determine whether we achieve similar positive results.

III. Novelty of Approach

The approach undertaken in this project is novel in several ways, primarily in how it combines sentiment analysis and topic modeling techniques to derive actionable insights from customer reviews. Existing solutions in sentiment analysis typically focus on either sentiment classification or clustering separately, often relying on basic Bag-of-Words or shallow feature extraction methods. However, the approach used here integrates both supervised and unsupervised learning techniques, providing a more comprehensive understanding of customer sentiment.

Combination of Sentiment Classification and Clustering

While many traditional sentiment analysis models focus on classifying reviews into positive or negative categories, this project introduces an additional layer of insight by applying unsupervised learning (LDA) to cluster reviews into dominant topics. This enables the extraction of not just sentiment but also key themes and customer pain points, which is often overlooked by conventional methods. By combining sentiment classification with topic modeling, the approach addresses the limitation of sentiment analysis systems that may fail to capture the underlying reasons behind customer opinions.

Advanced Feature Extraction with TF-IDF and CountVectorizer

The integration of both CountVectorizer and TF-IDF as feature extraction techniques allows for a more nuanced representation of the text. TF-IDF, in particular, helps emphasize important terms in the reviews, reducing the impact of frequently occurring, less informative words. This results in more accurate sentiment predictions compared to using CountVectorizer alone, which may not capture the full context or significance of specific terms. This dual approach enhances the model's robustness and improves sentiment classification accuracy, providing a more reliable sentiment analysis tool for e-commerce platforms.

Large-Scale Dataset

Many existing sentiment analysis models are trained on relatively small datasets, limiting their ability to generalize across a wide range of products and consumer behaviors. This project's use of the Amazon Electronics Product Review Dataset, with over 1.6 million entries, ensures that the models are trained on a diverse and large-scale dataset, enhancing their generalizability and scalability. This addresses the shortcoming of many existing solutions that may struggle to handle large volumes of data effectively.

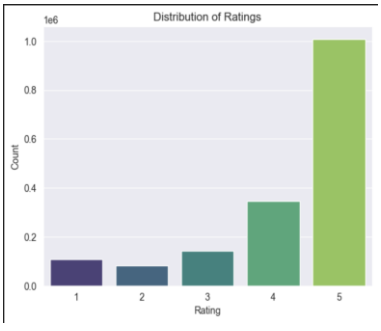
Evaluation Across Multiple Models

The project evaluates a variety of machine learning classifiers (Logistic Regression, SVM, Multinomial Naive Bayes, AdaBoost) on both CountVectorizer and TF-IDF features, providing a thorough comparison of their performance. This systematic evaluation enables the identification of the most effective model and feature combination for sentiment prediction, addressing the often-overlooked need for comprehensive model comparison in sentiment analysis.

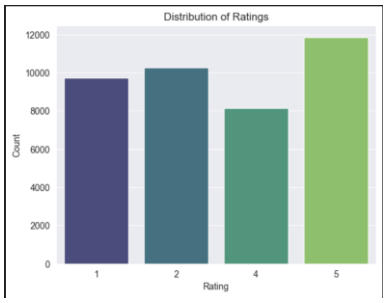
III. Dataset Analysis

A. Data Acquisition

For this project, we utilized the Amazon Product Review Dataset , available at Amazon Electronics Review Dataset, from the Electronics category, which contains over 1.6 million entries. We focused specifically on the reviewText and overall rating attributes, as these provided valuable insights for sentiment classification. The dataset, originally provided in JSON format, was processed to facilitate binary sentiment classification. To label the reviews, we classified those with overall ratings greater than 3 as positive (labeled as 1) and those with overall ratings less than 3 as negative (labeled as 0). Reviews with an overall rating of exactly 3 were excluded from the analysis, as they were considered neutral and did not contribute to the sentiment-focused analysis. This approach allowed us to concentrate solely on positive and negative sentiments, ensuring a more focused and relevant dataset for the subsequent analysis and processing stages.



(a) Original Data Distribution



(b) Balanced Data Distribution

Fig. 2: Data Distribution

B. Data Pre-processing

From Figure 2 (a), it can be seen that our original data distribution was highly imbalanced. The reviews with a rating of 5 were in large numbers compared to the other ratings, such as reviews with a rating of 1 or 2. This indicates that there were significantly more positive reviews than negative reviews in the original dataset. In order to analyze how the models behave in different data distributions, we carried out two approaches: one for the original data distribution and another for the balanced data distribution.

For the purpose of balancing the data, we performed selection of the top reviews from each sentiment category. Specifically, we selected the top 20,000 reviews for positive and negative sentiments based on a combined score derived from helpfulness, summary length, and review length. This was done to reduce the impact of data imbalance and focus on a more controlled dataset.

By selecting an equal number of reviews for both positive and negative sentiments, we aimed to create a more balanced dataset for analysis, as shown in Figure 2 (b). This approach was intended to allow for a better comparison of how models perform when trained on more balanced data compared to the original, imbalanced dataset.

For our data cleaning process, we perform the following tasks:

• **Tokenization**:
We break down the reviews into individual words, known as tokens, which are then used in the subsequent parsing process. The NLTK Python package is used to tokenize the reviews.

• **Removing Stop Words**:
Stop words are common words that do not contribute meaningful information to the text mining process and may negatively impact accuracy. Therefore, we remove these from our corpus. Using the NLTK package's built-in stop word list, we eliminate stop words, though we retain words like “not” and “no” to preserve the context in sentences such as “Not Good,” which would otherwise become “Good.”

• **Removing Hyperlinks**:
We observed that some reviews contained hyperlinks, which are irrelevant to our analysis. These are removed using the Beautiful Soup module.

• **Removing Punctuation**:
Punctuation marks are not useful for our analysis, so we remove them from the corpus using regular expressions.

• **Lemmatization**:
Lemmatization converts words to their base or root form, ensuring that our corpus is built with meaningful, consistent words. This process helps in improving the quality of the data for analysis.

C. Feature Extraction

(a) Features from the positive reviews

(b) Features from the negative reviews

Fig. 3: Review World Cloud

1) Bag-of-Words
The Bag-of-Words (BoW) approach is a technique for extracting features from text by capturing word occurrences. It focuses on the words and their frequencies within the text. Given the large volume of reviews in our dataset, considering all words as features would be computationally intensive. Therefore, we are selecting the top 2,000 most frequently used words to create our Bag-of-Words model.

For feature extraction, we are utilizing two main methods: **Count Vectorizer** and **TF-IDF**. When constructing the vocabulary for both methods, we set a minimum document frequency threshold of 5. This means that words appearing in fewer than 5 reviews are excluded. Additionally, we are considering both unigrams (single words) and bigrams (pairs of consecutive words) by setting the n-gram range to 1 and 2.

- *Count Vectorizer:* It converts a collection of text documents to a matrix of token counts. This implementation builds a sparse representation of the token occurrences.
- *TF-IDF:* Know as the term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. If a positive word occurs in a negative review multiple times or vice versa, the weight of such words are reduced in the TF-IDF representation.

IV. Classification

We define a series of classifiers following a similar approach used in *[9]*. The classifiers employed include Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Logistic Regression (LR), and AdaBoost using the Scikit-learn Python package. The classification metrics for all the classifiers are compared, and the best-performing classifier for our dataset is selected. This classifier is then evaluated in comparison with others to determine the most suitable approach for our dataset.

VI. Results

A. TF-IDF Accuracy Results:

The performance of the sentiment analysis models using TF-IDF (Term Frequency-Inverse Document Frequency) was evaluated at varying dataset sizes of 10k, 20k, 30k, and 60k reviews. The highest accuracy was achieved using Logistic Regression (TF-IDF), which reached 93.7% at the 60k dataset size, demonstrating strong performance in distinguishing positive and negative sentiment. The Support Vector Machine (SVM) model also performed well, with accuracy increasing from 87.1% at 10k reviews to 93.1% at 60k reviews. Multinomial Naive Bayes (TF-IDF) showed moderate performance, achieving 91.0% accuracy at 60k reviews, while AdaBoost (TF-IDF) achieved the lowest accuracy among the TF-IDF models, peaking at 89.4% at 60k reviews. These results indicate that TF-IDF, combined with Logistic Regression and SVM, is highly effective for sentiment classification on the Amazon Electronics dataset, with performance improving as the dataset size increases. The results are shown in Fig. 7

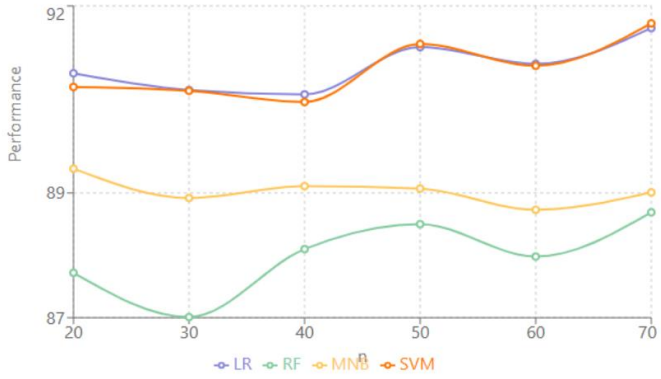


Fig. 7: Accuracy results for Balanced Data Distribution for TF-IDF

B. Count Vectorizer Accuracy Results:

In contrast, CountVectorizer was also evaluated for sentiment classification using various models, with accuracy results at dataset sizes of 10k, 20k, 30k, and 60k reviews. The Logistic Regression (CountVectorizer) model achieved the highest accuracy of 91.2% at 60k reviews, slightly underperforming compared to the TF-IDF version of the same model. The SVM (CountVectorizer) model performed similarly, with accuracy ranging from 85.5% at 10k reviews to 90.8% at 60k reviews. Multinomial Naive Bayes (CountVectorizer) demonstrated a steady performance, peaking at 89.2% accuracy at 60k reviews. AdaBoost (CountVectorizer) exhibited the lowest accuracy across all models, reaching a maximum of 87.3% at the 60k dataset size. Overall, while CountVectorizer performed well, it yielded slightly lower accuracy than TF-IDF, suggesting that TF-IDF may better capture the importance of words for sentiment classification in this context. The results are shown in Fig. 8

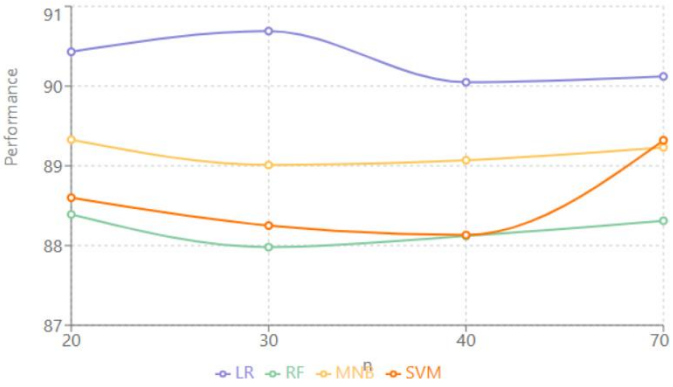


Fig. 8: Accuracy results for Balanced Data Distribution for Count Vectorizer

C. Final Model Results:

The final model, which used Logistic Regression with TF-IDF features, achieved a best cross-validation accuracy of 93.79%. The model was trained with the best parameters, including C = 1, max_iter = 100, and the saga solver. The classification performance on the testing data showed that the model was highly effective in sentiment classification, achieving an overall accuracy of 94%. Specifically, the model demonstrated strong performance for both classes (positive and negative), with precision of 0.95 for the negative class and 0.93 for the positive class. The recall values were also high, with the negative class achieving 0.93 and the positive class 0.95, meaning the model was good at identifying both positive and negative reviews. The F1-score for both classes was 0.94, indicating a balanced performance between precision and recall. The macro average and weighted average of precision, recall, and F1-score were all 0.94, reflecting the model's consistency across both classes. However, it is important to note that a ConvergenceWarning was raised during training, indicating that the model reached the maximum number of iterations without fully converging. This could potentially affect model stability, and further adjustments to parameters, such as increasing max_iter, may be beneficial. Despite this, the overall results demonstrate that the model performs very well in classifying reviews as either positive or negative.The confusion matrix is shown in Fig. 9

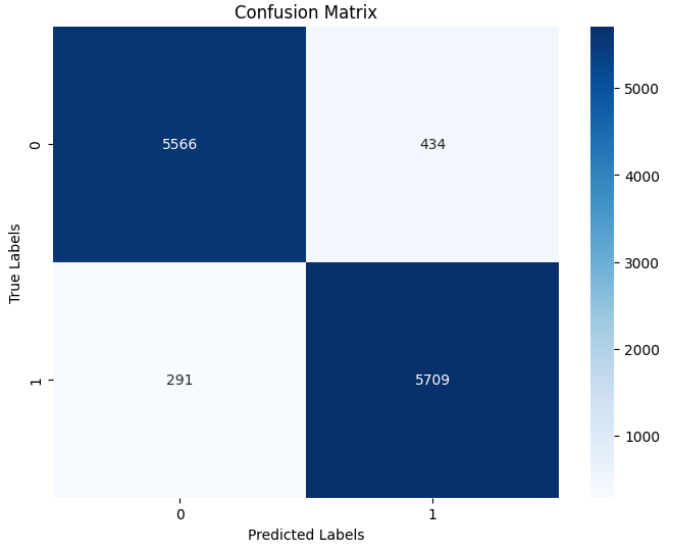


Fig. 9: Confusion Matrix

Fig.10: Training and Testing accuracy

In our research, we applied topic modeling techniques to analyze customer reviews and extract key themes related to specific products. Using Latent Dirichlet Allocation (LDA), we identified coherent clusters of reviews based on the most prominent topics discussed by customers.

After processing the reviews, we generated bigrams to capture meaningful word pairs, such as "battery life" or "screen resolution," which provided additional context beyond individual words. We then created a dictionary and corpus for the reviews, representing the text data in a format suitable for LDA modeling.

To determine the optimal number of topics, we trained LDA models with varying topic counts (ranging from 2 to 5) and evaluated their performance using coherence scores. The model with the highest coherence score was selected, ensuring that the identified topics were semantically meaningful and consistent.

Each review was assigned a dominant topic based on the LDA model's output, enabling us to group the reviews into clusters representing distinct themes. For each topic, we extracted the top representative words, such as "screen," "battery," or "sound," which highlighted the most frequently discussed aspects of the product.

Finally, we visualized the topic distributions and their relationships using interactive tools like pyLDAvis. This provided a clear understanding of the dominant themes, their relative prevalence, and the overlaps between topics. The results demonstrated how topic modeling can uncover valuable insights from customer feedback, offering actionable recommendations for product improvement and customer satisfaction. The results are shown in fig.11

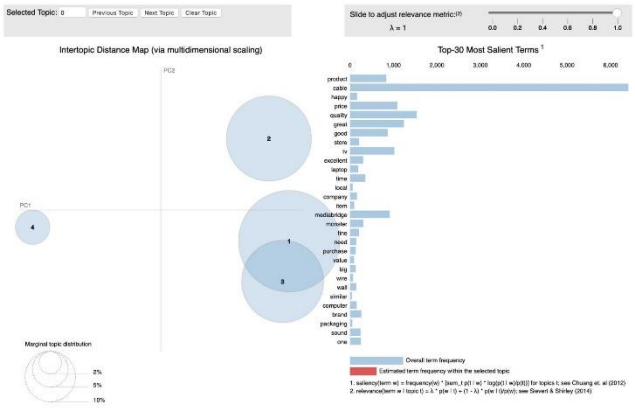


Fig. 11: Clustering Results

.

VII. CHALLENGES FACED:

One of the challenges we faced was that, after tuning parameters for the embedding model on the balanced data distribution, the results were nearly identical to the best classifier, SVM, in the Bag of Words (BoW) approach, even though we expected better performance. As a result, we decided to switch to the Tf-Idf for embedding, as it provided more useful features. Additionally, we observed that using traditional machine learning models on the original data distribution yielded better results than the BoW approach. We also tried applying Part-of-Speech (POS) tagging and selectively focusing on adjectives, as suggested in paper [5], but this showed little or no improvement. Given the large size of our dataset and limited hardware resources, we encountered significant difficulties in computing the results.

VIII. CONCLUSION

In this project, we applied a supervised learning approach to detect the polarity of reviews in our dataset, classifying them based on both balanced and original data distributions. After performing 5-fold cross-validation to evaluate our methods, we observed some interesting results. Logistic Regression (LR) emerged as the best-performing classifier, outperforming all other models in both the balanced and original data distributions. For the Bag of Words (BoW) approach, Support Vector Machine (SVM) with TF-IDF gave strong results, but Logistic Regression still held the highest performance. Additionally, the Multinomial Naive Bayes (MNB) classifier was fast in terms of computation and provided decent results, although not as high as Logistic Regression. We also found that the distribution of ratings had a significant impact on model performance, with the balanced data distribution yielding better results than the original one. Furthermore, we applied unsupervised learning techniques, such as K-Means and DBSCAN, for clustering analysis to group similar words and phrases in the reviews. These clusters revealed distinct patterns in customer sentiment, with positive and negative sentiments forming separate groups based on frequently used terms. The clustering also uncovered common themes across reviews, helping us understand the factors influencing customer satisfaction and dissatisfaction. Visualizations using t-SNE further illustrated the separation of these clusters, providing valuable insights into the key attributes that drive customer sentiment.

IX. Future Work

While the approach used in this project offers valuable insights, several limitations need to be acknowledged. One key limitation is the dependence on review ratings for sentiment classification. The ratings provided by customers may not always align with the sentiment expressed in the review text, which can introduce bias into the sentiment prediction process. For example, a review with a high rating may still contain negative sentiment or vice versa. Additionally, the current model may struggle with understanding nuanced or context-dependent sentiments, such as sarcasm or irony. Sentiment analysis models, particularly those based on traditional feature extraction methods like CountVectorizer and TF-IDF, may fail to capture the subtleties in customer feedback, leading to misclassification in complex reviews.

Another limitation is related to the use of Latent Dirichlet Allocation (LDA) for clustering reviews. While LDA is a powerful unsupervised technique for identifying topics, it has its constraints. The assumption

that reviews can be represented as a mixture of a fixed number of topics may not always hold true for the diverse and dynamic nature of customer feedback. Moreover, interpreting the topics generated by LDA can be challenging, as they may not always correspond directly to the themes that are most relevant to customer sentiment. Furthermore, while the approach has been tested on a large-scale dataset of 1.6 million reviews, scalability issues may arise when handling even larger datasets, particularly with computationally intensive techniques like LDA and training complex machine learning models. This could limit the real-time applicability of the system for businesses with rapidly growing review volumes

The approach is also specific to the Amazon Electronics category, and its performance may not generalize well to other industries or product categories. Future work could explore the use of deep learning models, such as Recurrent Neural Networks (RNNs) or Transformer-based models like BERT, which have shown superior performance in capturing contextual nuances and handling complexities like sarcasm and irony. These models could improve sentiment classification accuracy and offer a more robust understanding of customer reviews. Another avenue for future research involves incorporating multimodal data, combining textual reviews with numerical ratings and metadata (such as reviewer history and product features), to provide a more comprehensive view of customer sentiment.

In terms of topic modeling, dynamic techniques such as Non-negative Matrix Factorization (NMF) or Dynamic Topic Models (DTM) could replace LDA to capture evolving trends and themes in customer feedback. This would allow for more accurate detection of emerging topics and better insights into how customer sentiment shifts over time. Addressing data imbalance is another potential improvement; if future datasets have imbalanced sentiment distributions, methods like oversampling, undersampling, or class-weight adjustments could be implemented to enhance model performance.

Finally, for practical applications, future work could focus on enabling real-time sentiment analysis. By streamlining the data preprocessing, feature extraction, and classification steps, businesses could gain immediate insights from customer reviews as they are submitted. Additionally, the use of transfer learning could help the model generalize across different product categories, making it easier to apply sentiment analysis techniques to new datasets with minimal retraining. With these advancements, the model could become a more powerful tool for businesses to monitor customer sentiment and make informed decisions.

References

[1] James Barry. Sentiment analysis of online reviews using bag-of-words and lstm approaches. In *AICS*, 2017.

[2] Maria Soledad Elli and Yi-Fan Wang. Amazon reviews, business analytics with sentiment analysis.

[3] Xing Fang and Justin Zhan. Sentiment analysis using product review data. volume 2, page 5. Springer, 2015.

[4] Andrew Goldberg. Cs838-1 advanced nlp: Automatic summarization. Madison: University of Winsconsin-Madison, 2007.

[5] Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. Sentiment analysis on large scale amazon product reviews. In *Innovative Research and Development (ICIRD), 2018 IEEE International Conference on*, pages 1–6. IEEE, 2018.

[6] https://towardsdatascience.com/beyond-accuracy-precision-and recall. Evaluvation metrics definitions. www.towardsdatascience.com, 2016.

[7] Yi Sun Mingxiang Chen. Sentimental analysis with amazon review data. Stanford University, 2016.

[8] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[9] Qinxia Wang, X Wu, and Y Xu. Sentiment analysis of yelps ratings based on text reviews. 2016.