AutoRate: A Comparative Analysis of Discriminative and Generative Models for Review Ratings

Max Elgart
melgart@usc.edu

Rijul Raghu
jraghu@usc.edu

Anusha Poornesh
poornesh@usc.edu

Abstract

Customer reviews are an important source of information, influencing purchasing decisions and reflecting product quality and customer satisfaction. This report presents a comparative analysis of discriminative and generative models applied to Amazon product reviews, focusing on their performance in a 5-class and a 3-class rating system. We utilized a dataset comprising over 1.7 million Amazon reviews. Our findings show that discriminative models generally outperform generative models in accuracy and efficiency. The performance discrepancies between the models are discussed, highlighting the impact of model architecture and class complexity on classification accuracy. This project advances our understanding of the strengths and limitations of each model type in rating prediction for different class granularity.

1 Introduction

The ability to accurately predict scores for products from customer reviews presents an opportunity to enhance e-commerce experiences. There can be a discrepancy between the sentiment expressed in the text of a review and the numerical rating given. These reviews offer a rich dataset that is suited to exploring complex text classification problems.

Machine learning models, specifically those designed for text classification, can help leverage this vast dataset. These models are broadly categorized into two types: discriminative and generative. Discriminative models, such as Robustly Optimized BERT Approach (RoBERTa), focus on distinguishing between different classes based on the features derived from the input data. These models are optimized for precision and are highly effective in tasks that require direct classification.

Generative models, such as Text-to-Text Transfer Transformer (T5), learn the joint distribution of data and labels. This capability not only allows them to predict class labels but also to generate new text instances, providing a broader understanding of the data structure and the underlying language patterns.

This report delves into the performance of these model types in classifying Amazon product reviews across a 5-class rating system and a simplified 3-class system. It focuses on evaluating the performance of discriminative versus generative models, the applicability of generative models in tasks where classes are generated as the next token, and the influence of class complexity on model accuracy.

This analysis benefits from the diversity of the data, making it ideal for assessing model capabilities across different levels of classification granularity. By highlighting the strengths and limitations of each model type in handling different class configurations, this project aims to provide valuable insights in automated text analysis.

2 Related Work

Discriminative models have been widely documented for their robust performance in classification tasks and their ability to serve as a strong starting point for comparison. On the other hand, generative models are known for their potential to provide deeper insights into the nuances of language used [1]. There is not much research that explores the use of generative models for a highly classification-oriented task such as this; however, there are works that explore use of discriminative models for product and review rating predictions while also comparing performance across various class systems.

In a paper titled "NSL-BP: A Meta Classifier Model Based Prediction of Amazon Product Reviews", the authors introduce an approach to enhance the accuracy of Amazon product rating predictions [2]. They utilize a meta-classifier model that combines several machine learning techniques, including k-means clustering, LDA, Naïve Bayes, Logistic Regression, and SVM, into a two-level stacking ensemble model. Their results indicate that this combined approach outperforms each of the individual models in predicting product ratings.

In another paper titled "Amazon Books Rating prediction & Recommendation Model", the authors utilized the Amazon dataset to predict book ratings and build a recommendation system [3]. They employed various PySpark machine learning APIs for model development, and used cross validation and hyperparameter tuning for model generalization. The study revealed higher accuracy in binary classification compared to multiclass classification when predicting book ratings, showing the implications of classification approach choices in applications that focus on rating prediction.

In a study titled "Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT", the authors developed a prediction pipeline using BERT and T5 models for sentiment analysis on Amazon reviews, focusing on eco-friendly products [4]. The models achieved high accuracy (92% for BERT and 91% for T5), showing their effectiveness in detecting sentiments from customer reviews.

Another paper titled "Amazon Product review Sentiment Analysis using BERT" focuses on BERT for binary sentiment classification of Amazon product reviews [5]. It successfully implemented a model using the DistilBERT tokenizer and TFDistilBERT for sequence classification, achieving high accuracy rates (94.73% on training and 92% on validation).

Finally, a paper titled "Sentiment analysis classification system using hybrid BERT models" introduces a hybrid model combining BERT, BiLSTM, and BiGRU for enhanced sentiment analysis [6]. The author highlights the model's performance on sentiment classification tasks, where it outperformed other standard models by incorporating multi-feature fusion techniques, emphasizing its potential in extracting deeper context from texts.

3 Data

3.1 Data Description

In our project, we have utilized a portion of the dataset described in the paper, 'Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects' [7]. This dataset, originally compiled for extracting high-quality justifications from raw user reviews, is used in our project to train and test various language models.

The dataset for this project comprises of customer reviews for automotive products sourced from Amazon. It includes 1,710,653 reviews, and each review entry includes features such as star rating (1 to 5 stars), reviewer ID, product ID, review time, etc. For the purpose of this project, we will consider only the star rating associated with each review.

Along with the distribution of ratings for the 5-class classification task, we will combine stars in order to reduce the complexity of the data to a 3-class classification task. We do this by combining 1 and 2-star reviews to be a negative class, 3-star reviews as a neutral class, and 4 and 5-star reviews as a positive class.

3.2 Data Distribution

The dataset as seen in Table 1 shows a skew towards positive reviews in the 5-star rating system, with 5-star ratings comprising 72% of the total, 4-star at 14%, 3-star at 6%, 2-star and 1-star at 4% and 3%, respectively. The imbalance continues for the 3-class classification task because we combine ratings from the original dataset distribution. As seen in Table 2, there is a heavy skew of positive ratings (86%) compared to neutral (6%) and negative (6%) ratings. This imbalance highlights the challenge of overrepresentation of positive feedback in sentiment analysis tasks.

Rating	Number of Reviews
5-Star	1,233,434
4-Star	242,332
3-Star	102,626
2-Star	80,181
1-Star	52,080
Total Reviews	1,710,653

Table 1: Dataset distribution across the 5-class rating system

Sentiment	Number of Reviews
Positive	1,475,766
Neutral	102,626
Negative	132,261
Total Reviews	1,710,653

Table 2: Dataset distribution across the 3-class rating system

To address the skew in our dataset's distribution, we implement an undersampling strategy. This involves reducing the number of instances in the over-represented categories to match the number of reviews in the least represented categories in both tasks, 1-star and neutral categories. This ensures that our predictive model is not biased towards 5-star and positive ratings.

3.3 Data Preprocessing

The preprocessing of the dataset aims to transform raw text into a clean, standardized format suitable for analysis. Only the features required for the project, reviews, and their star ratings are retained, while the other features are removed. After dropping duplicated data and undersampling, it resulted in in 47,716 reviews per star rating in the 5-class task and 89,944 reviews per sentiment in the 3-class task.

Text normalization techniques, including converting all text to lowercase and removing all special characters, are applied to reduce the language's complexity and ensure consistency in token recognition. Contraction expansion is utilized to standardize text further, making it easier for the model to understand different forms of the same expression. Additionally, words are lemmatized to their root meaning. This allows less complex models, such as lstm and tf-idf, to recognize patterns within a given review text better.

3.4 Train, Validation and Test Split

After preprocessing the data, we split the dataset into 80% for training, 10% for validation, and 10% for testing.

4 Methodology

The methodology of our project is structured into four main phases: data processing, model selection, training, and evaluation. Having already explained data processing, the upcoming sections will detail model selection, the training procedures, and their comprehensive evaluation using key performance metrics.

4.1 Model Selection

Models were selected based on the need to explore and evaluate the performance of both discriminative and generative models.

4.1.1 Discriminative Models

The discriminative models used include:

- **TF-IDF with Logistic Regression**: This model serves as the baseline due to its straightforward methodology and is effective at providing a baseline understanding of linear relationships between textual features and review ratings. TF-IDF was used as it serves as a simple yet powerful tool for initial benchmarking.

- **Long Short-Term Memory (LSTM) as a Discriminative Model**: Using LSTM in a discriminative capacity allows the model to leverage its capability to remember long-term dependencies in text, making it suitable for understanding the context within customer reviews which is essential for accurate sentiment classification.

- **RoBERTa (Robustly Optimized BERT Approach)**: Chosen for its advanced ability to handle context better due to its enhanced training on more data and refined masking strategies. RoBERTa's encoder architecture helps in achieving better contextual understanding suitable for complex classification tasks.

4.1.2 Generative Models

The generative models used in this project include:

- **LSTM as a Generative Model**: LSTM can also generate text sequences based on the learned characteristics of the dataset, thus used here to predict review ratings as tokens which allows for a dynamic assessment of review sentiment.

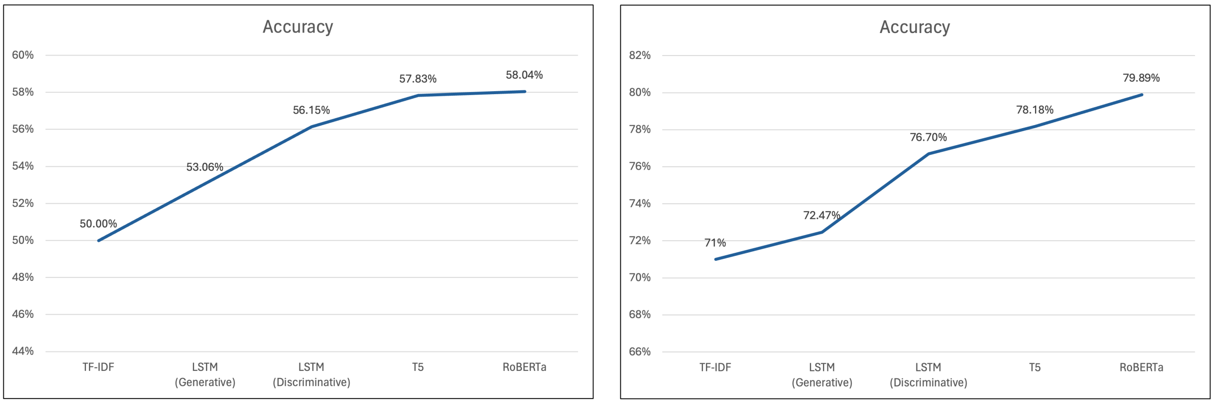
- **T5 (Text-to-Text Transfer Transformer)**: T5 is an encoder-decoder model that excels in tasks requiring a generation of textual output. It can be fine-tuned to focus specifically on generating ratings, making it ideal for tasks that require both understanding and generating text based on the input reviews.

Decoder-only models were not chosen for this analysis primarily because sequence generation with these models can be less predictable and harder to direct towards specific outcomes, such as star ratings. This makes them less suitable for the specific needs of predicting structured review ratings with high accuracy.

4.2 Training Methodology

Each model underwent a training process tailored to its specific architecture and operational requirements.

Figure 1: Comparison of Model Performance using Accuracy



(a) Performance Comparison for 5-Class System

(b) Performance Comparison for 3-Class System

4.2.1 Discriminative Models

Preprocessing: Standard preprocessing steps were applied, including text normalization and removal of irrelevant features.

Optimizer for Logistic Regression: Employed the Stochastic Gradient Descent optimizer to efficiently handle the large dataset.

Optimizer for LSTM and RoBERTa: Utilized the Adam optimizer, noted for its effectiveness in managing sparse gradients and addressing noisy problems in datasets.

4.2.2 Generative Models

Data Annotation: The training datasets for LSTM as a generative model and T5 were specially annotated to support sequence generation tasks. Each input was formatted with prefixes:

- “Review:” preceding the text of the review.
- “Rating:” before the target class label.

This format helps the models to learn generating the class label as the next token after the review content, aligning with the sequence-to-sequence learning approach.

Optimizers and Loss Management: Both LSTM (generative) and T5 were trained using the Adam optimizer. The primary training objective was to minimize loss across generated text sequences.

4.3 Evaluation Metrics

The performance of each model was evaluated using several key such as precision, recall, accuracy, and F1 score. These metrics were chosen to capture both the performance of each model in handling the multi-class classification task. The final model performance was then analyzed across various models and class granularity to determine how well each model met the project’s objectives.

5 Results

This section presents the results of each model. A comparison of the models’ accuracy can be seen in Figure 1.

5.1 TF-IDF with Logistic Regression

The Logistic Regression model achieved a test accuracy of 50% in the 5-class system and 71% in the 3-class system.

As the baseline model, it provides a fundamental comparison point across precision, recall and f1-score for evaluating the more advanced models, showing better results with fewer classes due to the reduced complexity in distinguishing between broader categories.

5.2 LSTM as a discriminative model

LSTM, used as a discriminative model, improved upon the baseline with accuracies of 56.15% in the 5-class system and 76.7% in the 3-class system.

The increase in accuracy can be attributed to LSTM’s ability to understand and process the temporal sequences inherent in text data. However, the model didn’t perform as well as RoBERTa or T5 as the architecture of these models are better at understanding nuanced language. LSTM is also trained from scratch as opposed to these models being pretrained on large datasets.

5.3 RoBERTa

RoBERTa further advanced the performance with a test accuracy of 58.04% for the 5-class classification and 79.89% for the 3-class classification, marking it as the top performer among the evaluated models.

Its encoder architecture and enhanced capabilities in contextual understanding and handling of complex linguistic structures contributed to its superior performance. RoBERTa is also pretrained on a lot more data than the LSTM and logistic regression models.

5.4 LSTM as a generative model

The generative LSTM achieved an accuracy of 53.06% for the 5-class rating system and 72.47% for the 3-class rating system. As expected, it outperforms the baseline TF-IDF for both systems as it actually captures sequentially dependencies and meaningful context.

However, it does not outperform the other models in either system. Given that this task is classification-oriented, the discriminative LSTM outperforms the generative LSTM too. It doesn't perform as well as RoBERTa and T5 due to their architectures being more adept at capturing nuanced language and their advantage of being pretrained on extensive datasets, unlike the LSTM, which is trained from scratch.

5.5 T5

T5 achieved an accuracy of 57.83% in the 5-class system and 78.18% in the 3-class system. Although slightly less effective than RoBERTa, T5 outperformed the other models.

Its encoder-decoder architecture allows for a flexible adaptation to varied text classification demands, making it highly suitable for tasks that require both a deep understanding of context. T5 is also pretrained on a huge corpus and hence has better at grasping language nuances.

However, RoBERTa performed better than T5 as it is specifically optimized for discriminative tasks. While T5 has broad generative capabilities, RoBERTa's training is highly focused on classification. T5 also requires more computation time as it generates sequences as part of its prediction process.

6 Discussion

This section provides a comprehensive analysis of the performance comparisons among the models evaluated.

6.1 Model Performance Comparison

From Figure 1, we observe a progression in accuracy starting from the simplest model, TF-IDF, moving to more complex models. The generative LSTM shows some improvement, followed by discriminative LSTM, and then T5, with RoBERTa emerging as the top performer. This trend highlights the increasing sophistication and capability of these models to handle the nuances of text classification, with discriminative models generally outperforming their generative counterparts.

6.1.1 Comparison of Top Performers: RoBERTa vs. T5

For a more detailed analysis, we compared the best-performing discriminative model, RoBERTa, with the top generative model, T5, using the same text dataset. This comparison helps to highlight the strengths and weaknesses of each approach in handling different classes of sentiment.

5-Class System: T5 demonstrated higher accuracy in classes 1 and 5, dealing effectively with strong negative and positive sentiments respectively, as illustrated in Figures 2(a) and 2(e).

RoBERTa showed better performance in classes 2, 3, and 4, which typically represent more nuanced sentiments that are less extreme. Figures 2(b), 2(c), and 2(d) display these distributions, and Table 3 lists examples where T5 outperformed RoBERTa for classes 1 and 5 and where RoBERTa outperformed T5 for classes 2, 3 and 4.

3-Class System: In 3-class classification, T5 was more accurate for class 1, effectively identifying negative sentiments as shown in Figure 3(a).

RoBERTa excelled in identifying class 2 sentiments, dealing better with neutral categories as detailed in Figure 3(b). For class 3, the performances of both the models are similar as seen in Figure 3(c). Table 4 lists examples where RoBERTa outperformed T5 and vice versa.

6.1.2 Explanation of Performance Differences

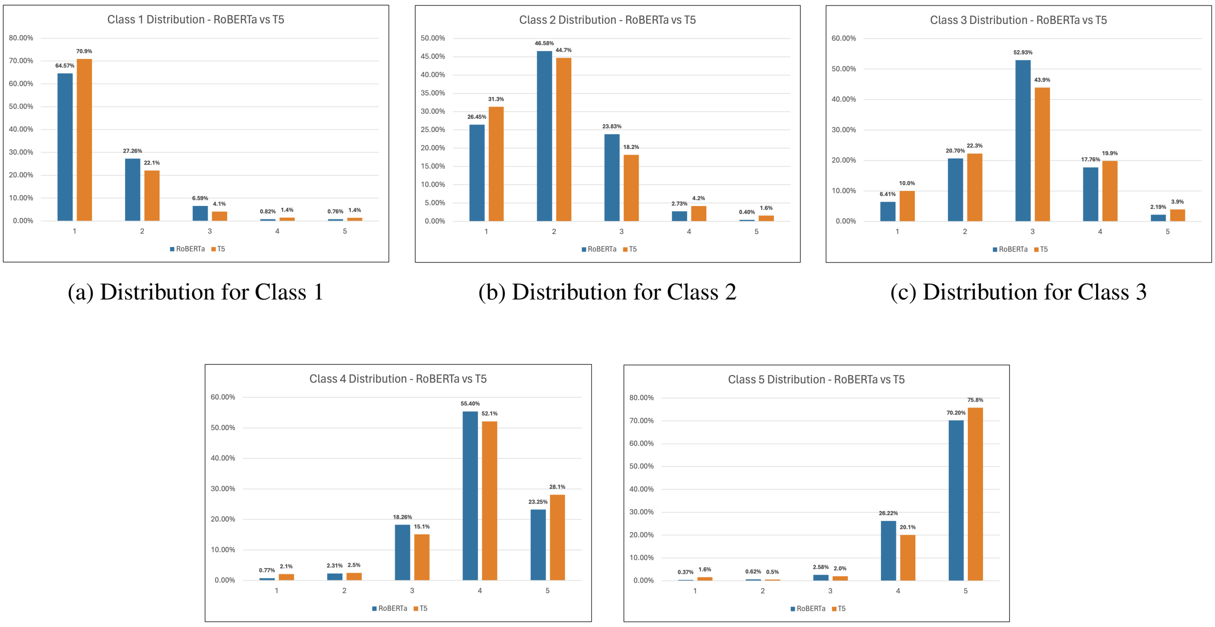
The observed differences in performance can be attributed to the inherent capabilities of each model:

T5: We can see from Table 3 and 4 that as a generative model, T5 excels in capturing the emotional

Table 3: Examples of Accuracy Contrast Between RoBERTa and T5 in Review Classification for 5-Class System

Review	True Rating	RoBERTa’s Prediction	T5’s Prediction
"cheap stuff"	1	2	1
"would not fit"	2	2	1
"it does the job"	3	3	4
"good tool to have works great"	4	4	5
"good product very informative"	5	4	5

Figure 2: Comparison of Distribution of Predicted Values for Actual Values of 5-Class System: RoBERTa vs. T5



(a) Distribution for Class 1

(b) Distribution for Class 2

(c) Distribution for Class 3

(d) Distribution for Class 4

(e) Distribution for Class 5

extremes of class 1 and 5 for 5-class systems and class 1 for 3-class systems, leveraging its ability to generate contextually relevant responses based on the training data. This might explain its strength in classes with more distinct, straightforward sentiment expressions.

RoBERTa: From Table 3 and 4 we can see that this model performs better on subtlety and context, essential attributes for classes 2, 3, and 4 in 5-class systems and class 2 in 3-class systems where sentiments are not as explicitly expressed. RoBERTa’s design, which focuses on understanding the deep connections between words, gives it an advantage in recognizing the subtle differences in the moderate classes.

6.2 Discriminative vs. Generative Models

The choice between discriminative and generative models in machine learning tasks is based on understanding their distinct characteristics, advantages, disadvantages, and the specific applications they are best suited for.

6.2.1 Discriminative Models

- **Approach to Learning:** These models are designed to identify the decision boundary between classes directly from input features. They do not model the underlying data distribution but focus on the relationship between the features and the labels.

- **Advantages:** These models have higher precision and are highly efficient during inference time.

- **Disadvantages:** They require large amounts of labeled data for optimal performance and also tend to perform less effectively on data that differs significantly from the training set.

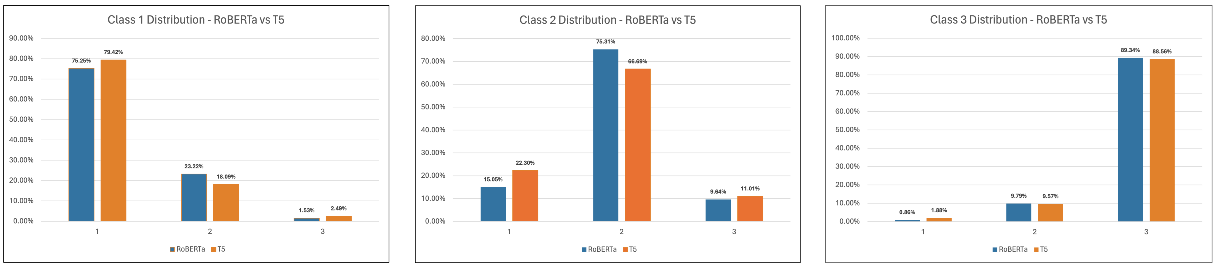
- **Output:** They directly predict the class labels in classification tasks.

- **Applications:** They are best suited for scenarios

Table 4: Examples of Accuracy Contrast Between RoBERTa and T5 in Review Classification for 3-Class System

Review	**True Rating**	**RoBERTa's Prediction**	**T5's Prediction**
"did not like the fit returned"	1	2	1
"its wrong size for my suburban"	1	2	1
"not that pleased"	2	2	1
"very big"	3	3	1
"high quality for very cheap"	3	3	2

Figure 3: Comparison of Distribution of Predicted Values for Actual Values of 3-Class System: RoBERTa vs. T5



(a) Distribution for Class 1

(b) Distribution for Class 2

(c) Distribution for Class 5

where precision and speed in classification are important and tasks where boundary clarity is essential.

6.2.2 Generative Models

- **Approach to Learning:** Generative models learn the joint probability distribution of the input features and the output. This allows them to not only predict output labels but also to model and generate new instances that are similar to the underlying data distribution.

- **Advantages:** These models excel in understanding the complete data space.

- **Disadvantages:** They often require significant computational resources making them slower in training and inference compared to discriminative models. Additionally, they can underperform on pure classification tasks.

- **Output:** They can predict class labels but are also capable of generating new data instances.

- **Applications:** They are best suited for scenarios where there is a need for creativity in output, such as in text generation, or in applications where training data is limited.

6.3 Classification via Generative Models

Generative models can transform traditional classification tasks into a sequence generation problem, where the class label is generated as the next token. This approach leverages the models' ability to reproduce contextually relevant text, making them efficient at nuanced classification tasks where context plays a pivotal role.

However, the effectiveness of generative models depends on well-annotated training datasets. They require significant computational resources, making them less practical for environments where processing power or data is limited.

6.4 Impact of Different Classes

Classification systems can vary widely in complexity based on the number of classes they encompass. For the purposes of this analysis, we have considered the performance of RoBERTa on 5-class and 3-class systems. The trend seen in RoBERTa's performance is similar to the trend in other models for different class granularity.

The complexity of a classification system impacts the difficulty level of modeling tasks. 5-class system introduces more opportunities for error. Table 5 highlights instances where the models tended to predict a rating either slightly higher or lower than the actual rating provided by users.

We examine the distribution of predicted values for each actual rating in both the 5-class and 3-class systems. Figures 4 and 5, respectively, provide visual representations of these distributions. In the 5-class system, as illustrated in Figure 4, the models often predict ratings that are adjacent to the true ratings for every class. This trend highlights

Table 5: Examples of reviews where predictions deviated by one rating point for 5-class System

Review	True Rating	Predicted Rating
"bad"	2	1
"do not waste your money"	2	1
"perfect replacement good quality"	4	5
"great price looks good no complaints"	4	5

Figure 4: Distribution of Predicted Values for Actual Values of 5-Class System

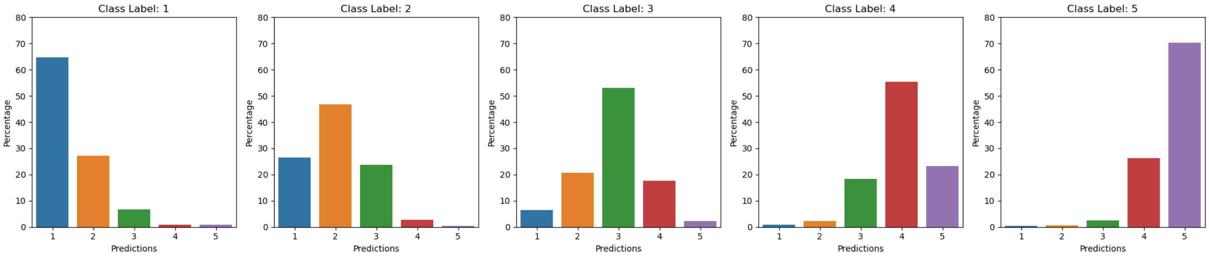
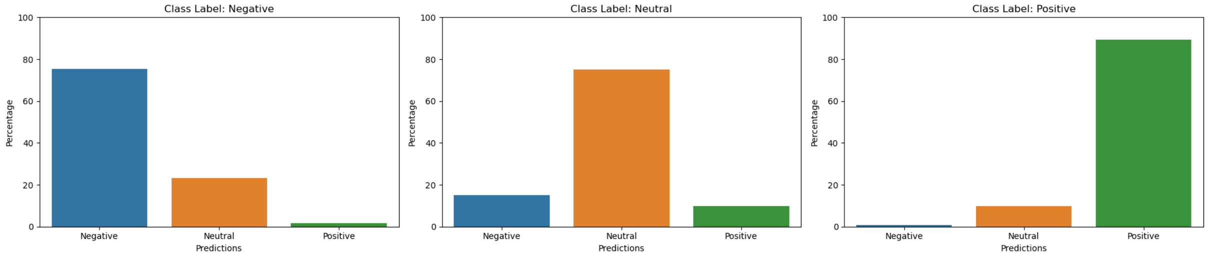


Figure 5: Distribution of Predicted Values for Actual Values of 3-Class System



the challenge of differentiating between nuances in customer sentiment.

The 3-class system, as shown in Figure 5, reduces this complexity by grouping the ratings into broader categories. As a result, models trained on the 3-class system often achieve higher accuracy and require a less nuanced understanding of the text.

7 Conclusion

Throughout this project, we have conducted an extensive analysis of discriminative and generative models, focusing on their capabilities within 5-class and 3-class classification systems.

Our findings showed that RoBERTa, a discriminative model, consistently excelled at handling nuanced sentiments within moderate categories, typically found in Classes 2, 3, and 4. T5, a generative model, demonstrated superior performance in identifying extreme sentiments of Classes 1 and 5. This distinction shows the complexity inherent in the 5-class system, where the need to discern fine gradations between classes posed significant challenges. In contrast, the 3-class system proved more robust for general sentiment analysis, simplifying the classification task and enhancing model accuracy by focusing on broader sentiment categories.

The implications of our study suggest avenues for future work, including refining the accuracy of generative models in moderate sentiment classes and exploring hybrid models that might combine the strengths of both discriminative and generative approaches.

In conclusion, the insights gained from this analysis not only advance our understanding of these models in sentiment analysis but also offer guidance on selecting and applying these models to meet specific analytical needs. This project shows the importance of model selection based on detailed performance analysis, ensuring that the chosen models align well with the specific challenges and requirements of sentiment classification tasks.

References

[1] Ruslan Salakhutdinov, "Learning deep generative models," *Annual Review of Statistics and Its Application*, vol. 2, pp. 361-385, 2015. https://www.cs.cmu.edu/~rsalakhu/papers/annrev.pdf

[2] Kumar, P., Dayal, M., Khari, M., Fenza, G., & Gallo, M. (2020). NSL-BP: "A Meta Classifier Model Based Prediction of Amazon Product Reviews". International Journal of Interactive Multimedia and Artificial Intelligence, 6(6), 95-103. https://doi.org/10.9781/ijimai.2020.10.001

[3] Lin, H.-P., Chauhan, S., Chauhan, Y., Chauhan, N., & Woo, J. (2023). "Amazon Books Rating prediction & Recommendation Model." https://arxiv.org/pdf/2310.03200

[4] Shaik Vadla, M. K., Suresh, M. A., Viswanathan, V. K. (2024). "Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT". Algorithms, 17(2). https://doi.org/10.3390/a17020059

[5] Inaniya, Y. (2023). "Sentiment Analysis using BERT: Amazon Review Sentiment Analysis." https://www.analyticsvidhya.com/blog/2021/06/amazon-product-review-sentiment-analysis-using-bert/

[6] Talaat, A.S. "Sentiment analysis classification system using hybrid BERT models". J Big Data 10, 110 (2023). https://doi.org/10.1186/s40537-023-00781-w

[7] Jianmo Ni, Jiacheng Li, and Julian McAuley, "Justifying recommendations using distantly-labeled reviews and fined-grained aspects," *Empirical Methods in Natural Language Processing (EMNLP)*, 2019. https://aclanthology.org/D19-1018.pdf

A APPENDIX

A.1 Classification Reports for 5-Class System

Table 6: Classification Report for TF-IDF

TF-IDF			
Class	**Precision**	**Recall**	**F1-Score**
1	0.56	0.62	0.59
2	0.40	0.36	0.38
3	0.42	0.41	0.41
4	0.45	0.42	0.43
5	0.64	0.70	0.67
Test Accuracy: 50%			

Table 7: Classification Report for LSTM (Gen)

LSTM (Generative)			
Class	**Precision**	**Recall**	**F1-Score**
1	0.62	0.62	0.62
2	0.43	0.42	0.42
3	0.44	0.44	0.44
4	0.49	0.48	0.48
5	0.66	0.70	0.68
Test Accuracy: 53.06%			

Table 8: Classification Report for LSTM (Disc)

LSTM (Discriminative)			
Class	**Precision**	**Recall**	**F1-Score**
1	0.65	0.62	0.64
2	0.46	0.48	0.47
3	0.49	0.45	0.47
4	0.52	0.53	0.52
5	0.69	0.71	0.70
Test Accuracy: 56.15%			

Table 9: Classification Report for T5

T5			
Class	**Precision**	**Recall**	**F1-Score**
1	0.60	0.75	0.67
2	0.50	0.37	0.44
3	0.53	0.47	0.50
4	0.53	0.44	0.53
5	0.70	0.74	0.73
Test Accuracy: 57.83%			

Table 10: Classification Report for RoBERTa

RoBERTa			
Class	**Precision**	**Recall**	**F1-Score**
1	0.66	0.65	0.65
2	0.48	0.47	0.47
3	0.51	0.53	0.52
4	0.53	0.55	0.54
5	0.73	0.70	0.72
Test Accuracy: 58.04%			

A.2 Classification Reports for 3-Class System

Table 11: Classification Report for TF-IDF

TF-IDF			
Class	**Precision**	**Recall**	**F1-Score**
1	0.72	0.71	0.71
2	0.62	0.61	0.62
3	0.79	0.79	0.79
Test Accuracy: 71%			

Table 12: Classification Report for LSTM (Gen)

LSTM (Generative)			
Class	**Precision**	**Recall**	**F1-Score**
1	0.74	0.70	0.72
2	0.63	0.65	0.64
3	0.81	0.83	0.82
Test Accuracy: 72.47%			

Table 13: Classification Report for LSTM (Disc)

LSTM (Discriminative)			
Class	**Precision**	**Recall**	**F1-Score**
1	0.75	0.78	0.77
2	0.68	0.67	0.68
3	0.87	0.85	0.86
Test Accuracy: 76.7%			

Table 14: Classification Report for T5

T5			
Class	**Precision**	**Recall**	**F1-Score**
1	0.76	0.80	0.78
2	0.71	0.66	0.69
3	0.87	0.88	0.88
Test Accuracy: 78.18%			

Table 15: Classification Report for RoBERTa

RoBERTa			
Class	**Precision**	**Recall**	**F1-Score**
1	0.82	0.75	0.79
2	0.70	0.75	0.72
3	0.89	0.89	0.89
Test Accuracy: 79.89%			

A.3 Predictive Performance Analysis for 5-Class System: RoBERTa vs. T5

Table 16: Distribution for Class 1

Actual Value: 1		
Class	**RoBERTa**	**T5**
1	64.57%	70.9%
2	27.26%	22.1%
3	6.59%	4.1%
4	0.82%	1.4%
5	0.76%	1.4%

Table 17: Distribution for Class 2

Actual Value: 2		
Class	**RoBERTa**	**T5**
1	26.45%	31.3%
2	46.58%	44.7%
3	23.83%	18.2%
4	2.73%	4.2%
5	0.40%	1.6%

Table 18: Distribution for Class 3

Actual Value: 3		
Class	**RoBERTa**	**T5**
1	6.41%	10.0%
2	20.70%	22.3%
3	52.93%	43.9%
4	17.76%	19.9%
5	2.19%	3.9%

Table 19: Distribution for Class 4

Actual Value: 4		
Class	**RoBERTa**	**T5**
1	0.77%	2.1%
2	2.31%	2.5%
3	18.26%	15.1%
4	55.40%	52.1%
5	23.25%	28.1%

Table 20: Distribution for Class 5

Actual Value: 5		
Class	**RoBERTa**	**T5**
1	0.37%	1.6%
2	0.62%	0.5%
3	2.58%	2.0%
4	26.22%	20.1%
5	70.20%	75.8%

A.4 Predictive Performance Analysis for 3-Class System: RoBERTa vs. T5

Table 21: Distribution for Class 1

Actual Value: 1		
Class	**RoBERTa**	**T5**
1	75.25%	79.42%
2	23.22%	18.09%
3	1.53%	2.49%

Table 22: Distribution for Class 2

Actual Value: 2		
Class	**RoBERTa**	**T5**
1	15.05%	22.30%
2	75.31%	66.69%
3	9.64%	11.01%

Table 23: Distribution for Class 3

Actual Value: 3		
Class	**RoBERTa**	**T5**
1	0.86%	1.88%
2	9.79%	9.57%
3	89.34%	88.56%