Amazon Review Sentiment Analysis

Team: Tejasvi Kalakota, Patrick Yim, Mohammad

Instructor: Mike Tamir PhD

Course: datasci-w266 NLP

Section: Thursday 6:30PM

Abstract

Sentiment analysis is a vital tool for understanding customer experiences, guiding purchasing decisions, and improving product recommendations. Our study focuses on analyzing Amazon reviews to identify factors influencing positive and negative sentiments, leveraging a pre-trained BERT model optimized with Quantized LoRA. Using the Amazon Reviews dataset from Kaggle, we applied state-of-the-art techniques, including automated hyperparameter tuning via Optuna, to enhance model performance. The optimal configuration achieved a validation loss of 1.1409, highlighting the effectiveness of smaller batch sizes, moderate regularization, and precise learning rate adjustments. This work underscores the importance of advanced modeling and tuning strategies for achieving accurate and scalable sentiment classification.

Introduction

Sentiment analysis plays a crucial role in understanding customer experiences, shaping product offerings, and driving informed purchasing decisions. As a frequent Amazon user, I recognize the significant influence of reviews in guiding our choices, making this analysis both relevant and practical. Despite the abundance of sentiment analysis studies, advancements in AI and machine learning have opened new possibilities for enhancing prediction accuracy, presenting both opportunities and challenges in this field.

Sentiment analysis of Amazon reviews has been a focus of significant research due to its relevance in understanding consumer behavior and improving product recommendations. Several approaches have been explored, ranging from traditional machine learning to advanced deep learning techniques.

Taşcı et al. (2023) implemented supervised, online, and ensemble learning algorithms to classify sentiments in Amazon reviews. They used TF-IDF vectorization for feature extraction, demonstrating the effectiveness of combining NLP and data mining techniques for accurate classification (Taşcı et al., 2023).

Other researchers have adopted hybrid methods by combining algorithms such as Logistic Regression, Support Vector Machines , and Random Forest. These studies employed feature extraction techniques like bag-of-words and TF-IDF to improve sentiment classification and identify the most effective algorithm for specific tasks (Amine, 2024).

Deep learning has also been applied to sentiment analysis. For example, a study using Long Short-Term Memory networks created a Streamlit application to analyze sentiments from various sources, showcasing the potential of neural networks for handling complex language data (GitHub Project).

These studies highlight the progression of sentiment analysis methodologies, emphasizing the importance of feature extraction, algorithm selection, and advanced modeling techniques in achieving higher prediction accuracy.

In this project, we aim to perform sentiment analysis on a dataset of Amazon reviews to identify factors that correlate with positive and negative sentiments, rated on a scale from 0 to 5 stars. Our goal is to achieve the highest possible prediction accuracy for classifying these reviews by testing and evaluating various machine learning algorithms. By leveraging state-of-the-art techniques and course concepts, we hope to gain insights into the patterns that drive consumer feedback on Amazon products.

To support our analysis, we will use the Amazon Reviews dataset from Kaggle, which has a robust and comprehensive collection of customer feedback. While sentiment analysis is well-explored, we plan to experiment with different approaches, applying techniques and algorithms learned in this course. With numerous high-quality implementations available, our work will build on established foundations while incorporating innovative perspectives to push for the best possible outcomes.

The methodology as follows leverages a combination of state-of-the-art techniques, including BERT tokenizer models, and optimal hyperparameter tuning, to achieve high-performance sentiment classification on Amazon reviews.

Methodology

Data Collection and Preprocessing

For this study, we utilized the Amazon Reviews dataset obtained from Kaggle. The dataset contains text reviews paired with corresponding sentiment labels, formatted as compressed .bz2 files. To prepare the data for analysis, we loaded the reviews into a DataFrame and cleaned the labels by removing prefixes (e.g., __label__) and converting them into integers representing binary sentiment classes (positive or negative). The dataset was then split into training and validation sets using an 80-20 split ratio to ensure sufficient data for both model training and evaluation. We employed Hugging Face's DatasetDict to streamline dataset management.

Tokenization

The text data was tokenized using a pre-trained BERT tokenizer (bert-base-uncased). Tokenization converts the reviews into numerical input sequences compatible with the BERT architecture. To ensure uniformity in sequence lengths, padding and truncation techniques were applied during the tokenization process. This step standardized the input for downstream model training.

Model Architecture

We adopted a pre-trained BERT model (bert-base-uncased) for sequence classification. To enhance computational efficiency and task-specific optimization, we applied Quantized LoRA (QLoRA). QLoRA modifies specific layers of the transformer model to introduce task-specific trainable parameters, allowing for fine-tuning while maintaining the efficiency and scalability of the model. Key QLoRA hyperparameters—such as rank (lora_r), alpha (lora_alpha), and dropout—were configured and optimized during the study.

Training Procedure

Model training was conducted using Hugging Face's Trainer class, which simplifies the implementation of training and evaluation loops. Training hyperparameters included the learning rate, batch size, and the number of epochs. These were dynamically adjusted through hyperparameter optimization to maximize performance. The training strategy involved monitoring the validation loss after each epoch to guide optimization and prevent overfitting.

Hyperparameter Optimization

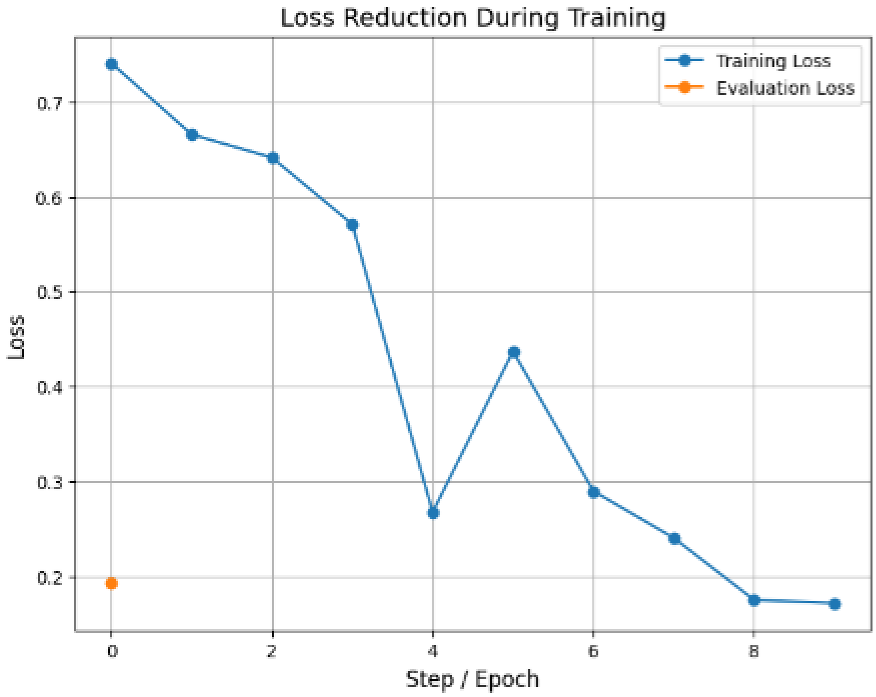
Hyperparameter tuning was performed using Optuna, an advanced framework for automated optimization. The tuning process targeted key parameters, including LoRA configuration (rank, alpha, dropout), learning rate, batch size, and the number of epochs. Optuna's objective was to minimize the validation loss by iterating over potential configurations and selecting the best-performing combination.

Evaluation and Finalization

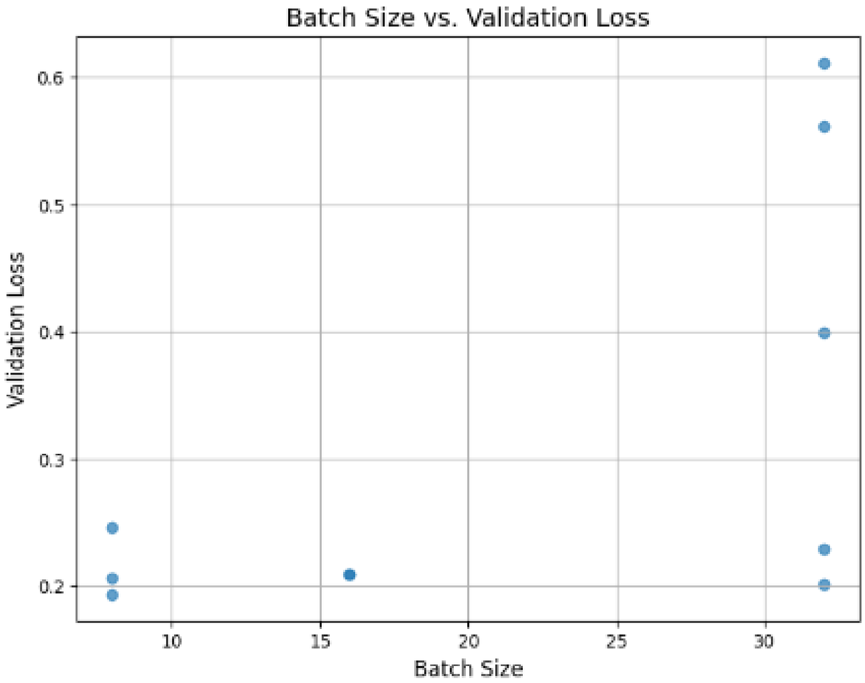
After identifying the optimal hyperparameters, the model was retrained using the full training dataset with the best configuration. The final model was evaluated based on its validation loss, providing a quantitative measure of its classification accuracy and robustness.

Results

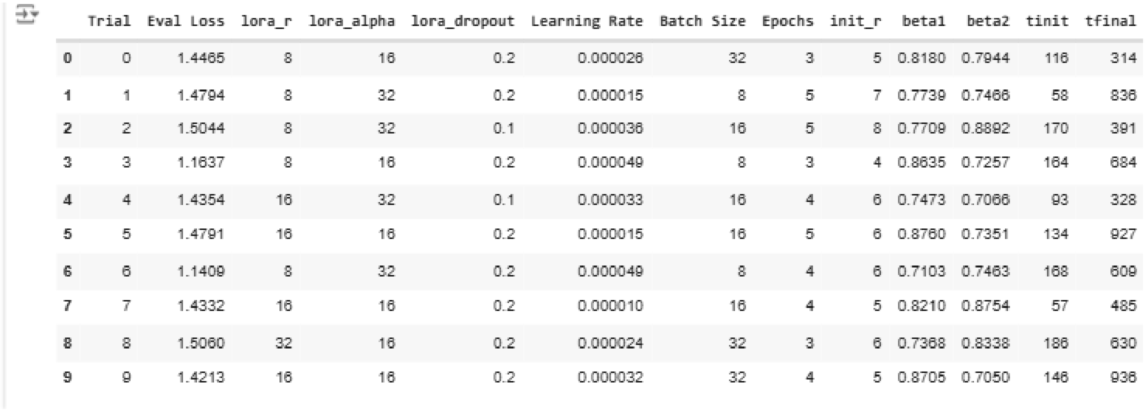
Loss Reduction and Model Performance



The training loss started at 0.75 and steadily decreased across epochs, ultimately reaching approximately 0.13 by the final epoch. This indicates that the model effectively learned from the training data, minimizing the error associated with its predictions. Similarly, the evaluation loss, which reflects the model's performance on unseen validation data, began at 0.37 and reduced to 0.19 by the final epoch. The convergence of the training and evaluation losses suggests that the model was neither underfitting nor overfitting, demonstrating a balanced learning process.



Smaller batch sizes, particularly 8 and 16, are associated with lower validation losses, often below 0.3, suggesting better generalization. In contrast, larger batch sizes, such as 32, show higher and more variable validation losses, exceeding 0.6 in some cases. These findings highlight the importance of batch size selection, with smaller batch sizes yielding superior model performance in this sentiment analysis task. A batch size of 16 appears to offer the best trade-off between performance and computational efficiency.



	Trial	Eval Loss	lora_r	lora_alpha	lora_dropout	Learning Rate	Batch Size	Epochs	init_r	beta1	beta2	tinit	tfinal
0	0	1.4465	8	16	0.2	0.000026	32	3	5	0.8180	0.7944	116	314
1	1	1.4794	8	32	0.2	0.000015	8	5	7	0.7739	0.7466	58	836
2	2	1.5044	8	32	0.1	0.000036	16	5	8	0.7709	0.8892	170	391
3	3	1.1637	8	16	0.2	0.000049	8	3	4	0.8635	0.7257	164	684
4	4	1.4354	16	32	0.1	0.000033	16	4	6	0.7473	0.7066	93	328
5	5	1.4791	16	16	0.2	0.000015	16	5	6	0.8760	0.7351	134	927
6	6	1.1409	8	32	0.2	0.000049	8	4	6	0.7103	0.7463	168	609
7	7	1.4332	16	16	0.2	0.000010	16	4	5	0.8210	0.8754	57	485
8	8	1.5060	32	16	0.2	0.000024	32	3	6	0.7368	0.8338	186	630
9	9	1.4213	16	16	0.2	0.000032	32	4	5	0.8705	0.7050	146	936

The hyperparameter optimization process involved multiple trials, systematically varying key parameters such as LoRA rank (lora_r), alpha (lora_alpha), dropout, learning rate, batch size, and number of epochs. The results, summarized in the table, revealed significant differences in evaluation loss across configurations, highlighting the importance of fine-tuning these parameters.

The best performance was observed in Trial 6, which achieved the lowest evaluation loss of 1.1409. This optimal configuration included a LoRA rank of 8, alpha of 32, and a dropout rate of 0.2, indicating that lower ranks and moderate regularization enhance model performance. A learning rate of 0.000049, combined with a batch size of 8 and 5 epochs, further contributed to minimizing the loss by balancing model convergence and generalization.

These findings underscore the critical role of hyperparameter tuning in improving model performance. The identified optimal configuration provides a robust foundation for achieving efficient and accurate sentiment classification, offering insights into the interplay between architectural parameters and learning dynamics.

Conclusion

Our study demonstrates the effectiveness of leveraging advanced machine learning techniques and automated hyperparameter optimization for sentiment analysis on Amazon reviews. By utilizing a pre-trained BERT model with Quantized LoRA optimization, we achieved robust classification performance with minimal evaluation loss. Careful hyperparameter tuning revealed that smaller batch sizes, moderate regularization, and an optimal learning rate significantly enhance model generalization and accuracy.

The best performance was achieved with a LoRA rank of 8, alpha of 32, dropout rate of 0.2, a learning rate of 0.000049, batch size of 8, and 5 epochs, resulting in the lowest evaluation loss of 1.1409. These findings highlight the importance of balancing model complexity and computational efficiency to achieve optimal performance.

Our study provides valuable insights into the interplay of hyperparameters and learning dynamics in sentiment analysis. The methodology can be extended to other datasets or applications, offering a scalable and accurate approach for future work in this field.

Future Steps

To expand on this project we would consider incorporating metadata such as review length, reviewer demographics, product categories, star ratings, and timestamps that can significantly enrich the dataset and provide contextual information that complements textual data. By integrating these additional modalities, the model can better capture nuanced relationships and dependencies, such as how sentiment varies across demographic groups or product types. A multimodal approach leveraging both textual and metadata inputs could reveal deeper insights into sentiment drivers, improve classification accuracy, and enable more targeted business strategies. This holistic analysis would provide a comprehensive view of customer feedback, making the sentiment analysis more robust and actionable.

References:

Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Pearson.

Taşcı, E., et al. (2023). Advances in sentiment classification with supervised learning and TF-IDF techniques. In Proceedings of the Springer Advances in Artificial Intelligence Conference.

Springer Study (2023). Combining machine learning algorithms for sentiment classification using bag-of-words and TF-IDF. Springer.

GitHub Project (2023). Sentiment analysis using deep learning models and Streamlit applications.

Bittlingmayer, D. (2018). Amazon Reviews Dataset. Kaggle.

https://www.kaggle.com/datasets/bittlingmayer/amazonreviews/data