Contents

Enhancing Sentiment Analysis of Yelp Restaurant Reviews with a RoBERTa-GRU Hybrid Model

1 Introduction

1.1 Background

Sentiment analysis has evolved significantly, transitioning from basic text classification methods to advanced deep learning techniques. This field's growth is particularly marked by the advent of social media platforms, where vast amounts of opinionated text data are generated daily. The analysis of such data, especially in platforms like Yelp, is crucial for businesses and analytics, as it provides insights into customer preferences and feedback.

1.2 State of the Art

Recent advancements in sentiment analysis have been driven by deep learning models, which offer superior capabilities in handling natural language's complexity. Transformer models, particularly the Bidirectional Encoder Representations from Transformers (BERT) and its variants like RoBERTa, have set new benchmarks in the field. However, these models often require enhancements to address specific challenges like long-range dependencies and contextual nuances in text. Hybrid models combining transformers with recurrent structures like GRUs have emerged as a solution, offering the benefits of both architectures

1.3 Problem Statement

Despite these advancements, sentiment analysis of Yelp reviews presents unique challenges. The class imbalance in such datasets and the intricate nature of user-generated content necessitate a more nuanced approach. Traditional models often fail to capture the subtleties of sentiment, leading to less accurate or biased interpretations.

1.4 Proposed Solution and Objectives

This report introduces a RoBERTa-GRU hybrid model specifically designed for the sentiment analysis of Yelp restaurant reviews. The objectives of this study are threefold: to develop a model that accurately classifies sentiments in Yelp reviews, to address the challenges posed by class imbalances and contextual complexities, and to analyze and compare results of different approaches to training the model.

1.5 Scope of the Project

The scope of this project encompasses a thorough review of related works, development and implementation of the RoBERTa-GRU model, extensive testing and validation of the model's performance, and a critical analysis of the results in comparison with state-of-the-art methods.

This study focuses on the Yelp restaurant review dataset, aiming to classify reviews into sentiment categories accurately. The scope includes model development, data preprocessing, training, and evaluation, with an emphasis on overcoming the challenges posed by imbalanced datasets in sentiment analysis.

2 Related Works

This section surveys contemporary methodologies in sentiment analysis, categorizing them into two primary streams: machine learning and deep learning approaches.

2.1 Machine Learning Approaches

Traditional machine learning techniques have been foundational in sentiment analysis. A study by Hemakala and Santhoshkumar (2018)[1] analyzed Indian Airlines feedback using classical algorithms such as Decision Trees and Support Vector Machines, among others. Their findings revealed AdaBoost as the most precise model. Similarly, Makhmudah et al. (2019)[2] applied SVM to a dataset of tweets about homosexuality in Indonesia, achieving high accuracy after preprocessing steps like lemmatization and TF-IDF feature extraction.

Further research includes Alsalman's (2020)[3] analysis of Arabic tweets using Multinomial Naive Bayes, which underscored the effectiveness of advanced tokenization and TF-IDF in sentiment analysis. Tariyal et al[]. (2018) compared various algorithms for product review tweets, highlighting the superiority of CART in terms of accuracy. Gupta et al. (2019) and Jemai et al[4]. (2021) extended this exploration to include an array of machine learning models, demonstrating a range of accuracies across different datasets, including Sentiment140 and Twitter samples.

2.2 Deep Learning Approaches

The advent of deep learning has ushered in more sophisticated sentiment analysis models. Ramadhani and Goo (2017) implemented a Multilayer Perceptron (MLP) for analyzing a bilingual tweet dataset, while Demirci et al. (2019) focused on Turkish tweets using a similar approach. These studies highlighted the significance of thorough data preprocessing and the use of Word2vec embeddings in enhancing model performance.

Innovations in hybrid models combining CNNs with LSTM or Bi-LSTM have also shown promise. Rhanoui et al. (2019)[5] and Tyagi et al. (2020)[6] employed such models for analyzing various text sources, from news articles to tweets, demonstrating their efficacy in capturing complex sentiment nuances. Notably, the integration of attention mechanisms, as explored by Jang et al. (2020)[7], further improved model accuracy in analyzing movie reviews.

Hossain et al. (2020)[8] and Yang (2018) further expanded on hybrid architectures, combining CNNs with LSTM and RNN, respectively, for diverse applications, from restaurant reviews to general sentiment analysis. Harjule et al. (2020)[9] compared machine learning and deep learning methods, revealing the strengths of LSTM in handling complex datasets.

To summarize, while traditional machine learning methods have laid the groundwork in sentiment analysis, deep learning, particularly hybrid models and advanced embedding techniques, have significantly advanced the field, offering more nuanced and contextually aware sentiment predictions.

3 Methodology

3.1 RoBERTa-GRU Model Architecture

3.1.1 RoBERTa

The model integrates the RoBERTa layer as the foundational encoding mechanism. RoBERTa stands as an enhancement of the BERT architecture, which itself is rooted in the Transformer framework, known for its prowess in sequence-to-sequence predictive tasks. The Transformer's innovative self-attention mechanism is leveraged over traditional recurrent or convolutional methods, offering a more nuanced approach to identifying relationships within the input data. This mechanism is adept at assigning significance to related pieces of information, thereby streamlining sequences for optimized processing.

In essence, the Transformer architecture is composed of two principal segments: an encoder that processes the input text and a decoder that generates output based on that processed information. However, our model specifically harnesses only the encoder segment provided by RoBERTa for text encoding purposes.

The original BERT model was a pioneering force in mitigating the constraints of directional language processing, introducing capabilities such as Masked Language Modeling and Next Sentence Prediction to enhance context awareness. Unlike BERT's one-time static masking, RoBERTa introduces a dynamic masking technique that varies the masked tokens across different training epochs, thus enriching the learning context for the model.

RoBERTa's efficiency is further increased by its use of byte-level Byte Pair Encoding, which streamlines the tokenization process with a more compact vocabulary, reducing computational load relative to BERT's character-level approach.

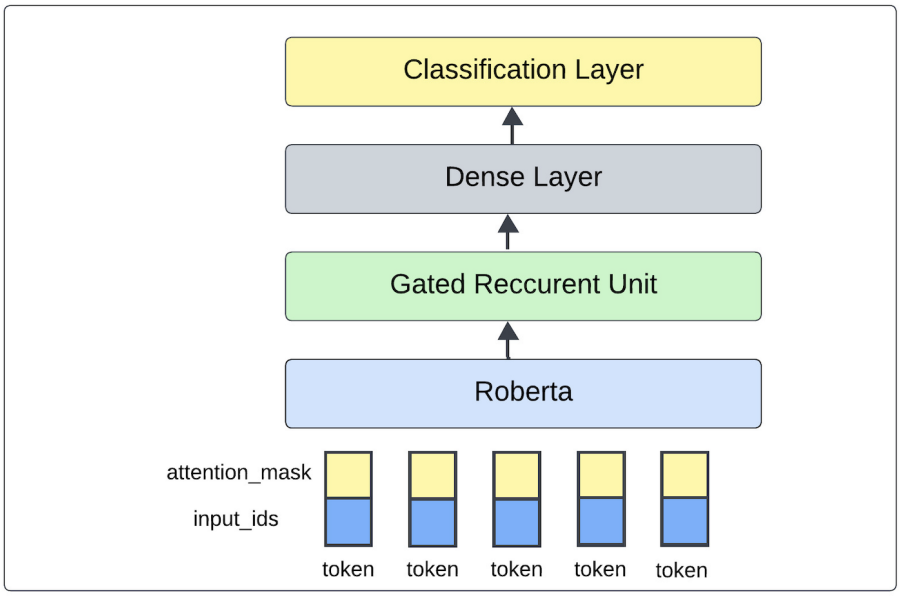


Figure 1: The architecture of the proposed RoBERTa-GRU model for Yelp Review Sentiment Analysis.

In practice, the RoBERTa tokenizer is utilized to deconstruct the textual data into subword units, or tokens, drawing upon its extensive pretraining to preserve the semantic essence of the language. Each token is mapped to a unique identifier within RoBERTa's vocabulary, complemented by attention masks that highlight the token's significance in the dataset. Consequently, these tokens, with their assigned input IDs and attention masks, are processed through RoBERTa's multi-layered neural network, each layer containing 768 hidden units, to capture a rich, contextual representation of the input text.

To enhance the interpretive power of the model, a Gated Recurrent Unit (GRU) is deployed post-RoBERTa encoding. The GRU's role is to synthesize the temporal relationships inherent in the tokenized data, allowing the model to generate predictions with a higher degree of accuracy, benefiting from the compounded contextual insights of RoBERTa and the sequential processing strengths of the GRU.

3.1.2 Gated Recurrent Unit (GRU)

The GRU is an advanced recurrent neural network designed to enhance the model's capacity for processing sequences by resolving the vanishing gradient issue commonly encountered in standard RNNs. Its architecture is composed of two gates: the update gate and the reset gate, which collaboratively regulate the flow of information through the network.

The update gate's role is to determine the degree to which the unit retains the previous state, while the reset gate influences the extent to which new input is allowed to modify the memory content. These gates operate in tandem to dynamically manage memory, effectively capturing dependencies for varying sequence lengths.

Mathematically, the operations within a GRU at any given time step t can be described by the following equations[10]:

$$
\begin{aligned}
\mathrm{z}_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
\mathrm{r}_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
\tilde{h}_t &= \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \\
\mathrm{h}_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
\end{aligned}
$$

Here, σ denotes the sigmoid activation function, which ensures the gates' outputs range from 0 to 1, and $\odot$ represents element-wise multiplication. W and U are the weights, while b denotes bias terms for their respective gates. The values z_t, r_t, and $\tilde{h}_t$ correspond to the update gate, reset gate, and candidate activation at time t, respectively. The final hidden state h_t is a weighted combination of the previous hidden state and the candidate activation, allowing the GRU to effectively retain or discard information across time steps, enhancing its ability to model long-distance relationships within the data.

The output from the GRU is then flattened, preparing it for subsequent dense layers and facilitating the end-to-end sentiment classification process.

3.1.3 Dense Layer and Classification

The dense layer of the RoBERTa-GRU model architecture plays a crucial role by connecting the feature-rich representations obtained by the GRU with the final classification judgments. The primary purpose of the dense layer is to compress the high-dimensional output generated by the GRU into a more sophisticated space, where each dimension is associated with possible emotion categories.

During the implementation, the forward pass begins with the RoBERTa model producing the final hidden state, which serves as the sequence output. The output is next fed into a GRU layer, and the ultimate hidden state of the GRU sequence is chosen for further processing.

Batch normalization is selectively done to maintain the model's robustness across various batch sizes. Stabilizing the learning process and normalizing the feature representation is essential in order to improve generalization across various inputs.

After the process of normalization, dropout is used as a means of regularization in order to mitigate the problem of overfitting. Dropout, by selectively deactivating a portion of input units, promotes the acquisition of resilient characteristics by the model that do not depend on particular activations from the preceding layer.

The output of the dropout layer is then transformed into a flattened format. The flattening procedure is important because the dense layer requires a one-dimensional vector for each input in the batch, and this method reshapes the data properly.

After being flattened, the output is sent into the first dense layer, where a Rectified Linear Unit (ReLU) activation function is used. This function provides non-linearity, enabling the model to acquire intricate relationships between features and classes. The next compact layer, often known as the classification layer, generates the logits, which are the unprocessed and unnormalized scores for each sentiment category.

Ultimately, these logits may be converted into probabilities using a softmax function outside the model's forward pass. This conversion is usually done via the loss function during training or directly used when generating predictions during inference. The softmax function guarantees that the output probabilities add up to one and are directly proportional to the model's confidence in each class, so successfully ending the sentiment analysis process.

3.2 Model Training Parameters

The training of the sentiment analysis model is governed by carefully selected parameters to ensure optimal learning and performance on the Yelp review dataset.

3.2.1 Categorical Cross-Entropy Loss Function

The categorical cross-entropy loss, also known as softmax loss, is pivotal in multi-class classification tasks. It quantifies the difference between the predicted probability distribution generated by the model and the true distribution of the labels. Mathematically, for a single sample with a true label y in one-hot encoded form and a predicted probability distribution $\hat{y}$, the loss can be expressed as:

$$L(y, \hat{y}) = -\sum_{j} y_j \log(\hat{y}_j) \tag{1}$$

Here, y_j is the binary indicator of the class label j and $\hat{y}_j$ is the predicted probability of the class label j. The logarithmic component penalizes incorrect predictions with a higher value, emphasizing the correct classification in the model's objective.

3.2.2 AdamW Optimizer

The AdamW optimizer is an extension of the traditional Adam optimizer, introducing an explicit decoupling of the weight decay from the gradient updates. It retains Adam's benefits of adaptive learning rate methods and combines them with a more robust weight regularization technique. The optimizer updates the weights iteratively based on the first (m_t) and second (v_t) moment estimates of the gradients:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{2}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{3}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t - \alpha\lambda\theta_t \tag{4}$$

In these equations, g_t is the gradient of the loss function with respect to the weights at time step t, α is the step size or learning rate, λ is the weight decay term, ϵ is a small scalar added to improve numerical stability, and θ_t represents the model parameters at time step t. The hyperparameters β_1 and β_2 control the exponential decay rates of the moving averages, with common default values being 0.9 and 0.999, respectively.

4 Dataset

4.1 Training Dataset

The dataset utilized for training and evaluating the RoBERTa-GRU model comprises Yelp reviews, which are inherently imbalanced with respect to sentiment labels. Initial data distribution, as shown in the Figure 3 below, indicates a majority of 'Positive' sentiments compared to 'Negative' and 'Neutral'.

4.2 Data Preprocessing

The preprocessing of Yelp review data is a crucial step to ensure its suitability for analysis. This process involves cleaning and normalizing the raw text. Key steps include the removal of irrelevant elements like URLs, stop words, and special characters, and standardizing text through case folding. This preprocessing helps in reducing noise and enhances the quality of data fed into the model. For the training dataset used in this study, the following preprocessing functions were applied:

- **Text Normalization:** Converting all text to a standard format, typically lower case, to ensure consistency.
- **Stop Words Removal:** Eliminating common words that provide little value in understanding the sentiment of the text.
- **Punctuation Removal:** Stripping punctuation marks, which are generally not informative for sentiment analysis models.
- **Data Augmentation:** Applying techniques such as back-translation to enhance the robustness of the dataset by artificially increasing its size and variety.
- **Tokenization:** Splitting text into individual words or tokens to facilitate further processing.

Each of these steps plays a pivotal role in cleaning and preparing the dataset for effective model training. By streamlining the data into a format that the RoBERTa-GRU model can efficiently process, we enhance the model's capacity to capture and learn from the underlying sentiment expressed in the Yelp reviews.

4.3 Class Distribution in Training Dataset

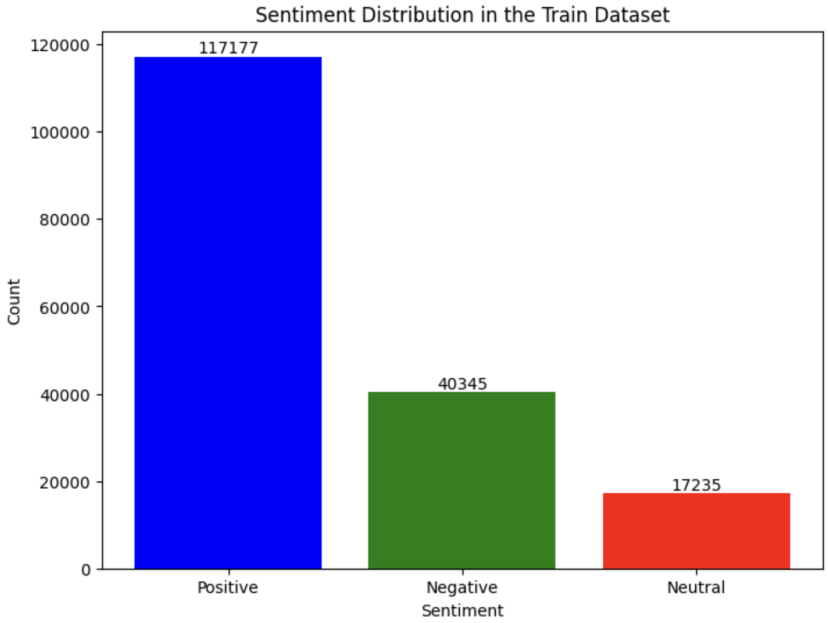


Figure 2: Initial distribution of sentiment classes in the Yelp review dataset.

Given the imbalance, data augmentation techniques were employed to balance the representation across classes, aiming for a uniform distribution of 50,000 samples per class. This approach aligns with practices observed in other studies, where datasets such as the Twitter US Airline dataset were augmented using GloVe word embeddings to achieve a balanced class distribution for robust model training and evaluation.

4.3.1 Data Augmentation Process

The data augmentation process for the Yelp review dataset involved generating synthetic samples to balance the dataset. Techniques such as synonym replacement and back translation were employed to enhance the minority 'Negative' and 'Neutral' classes without losing the semantic integrity of the reviews. The goal was to provide a dataset conducive to training a model that could generalize well across all sentiment classes.

4.3.2 Post-Augmentation Dataset

After augmentation, the Yelp review dataset achieved an equal distribution of sentiments, ensuring that each class had an equal chance of being correctly identified by the RoBERTa-GRU model, thereby mitigating potential biases during the model's training phase.

4.4 Test Dataset

While augmentation addressed class imbalances in the training dataset, the test dataset presents its own imbalances. As illustrated below, the distribution of sentiment classes within the test set is not uniform, with a predominance of positive reviews compared to negative and neutral ones.

This inherent skewness is critical to consider when analyzing the model's performance, as it can affect metrics such as accuracy and F1 score. A model's ability to correctly predict minority classes is as important as its overall accuracy, especially in practical applications where all sentiment categories are equally important for comprehensive analysis. Therefore, results on the test set will be examined with additional metrics that account for this imbalance, such as precision, recall, and the confusion matrix, to ensure a thorough evaluation of the model's predictive capabilities across all classes.

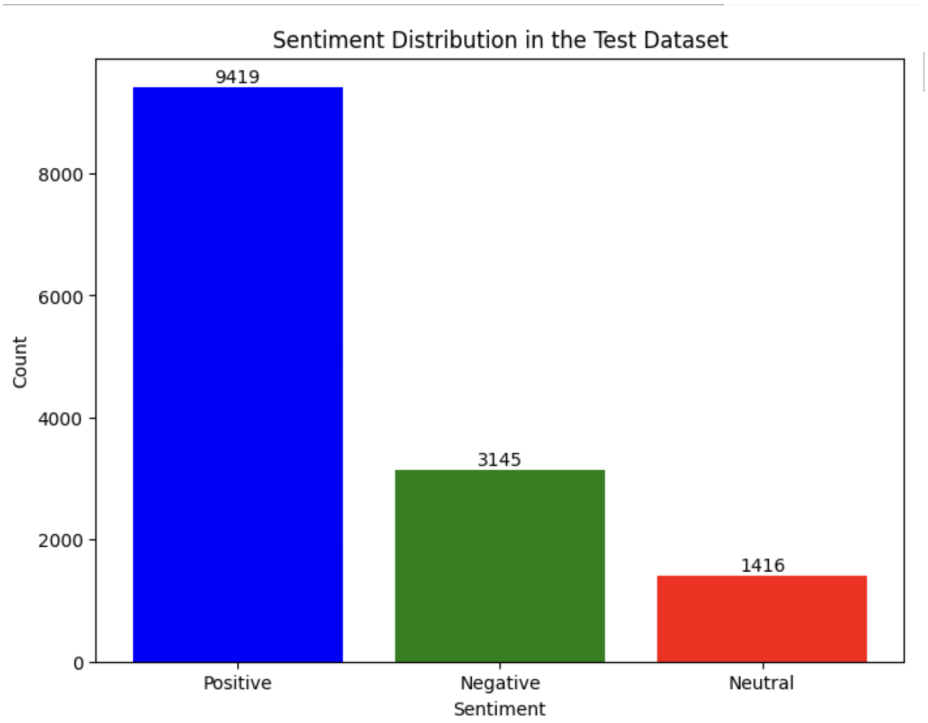


Figure 3: Test Dataset Class Distribution

5 Hyperparameter Tuning

The process of hyperparameter tuning was crucial in optimizing the performance of the RoBERTa-GRU model. A grid search approach was employed to determine the optimal settings. Key hyperparameters, including the epochs, training batch size, and learning rate, were varied to assess their impact on the model's effectiveness. The tuning was done on a subset of the entire dataset, which had 15,000 data instances for each class.

5.1 Hyperparameter Tuning Analysis

The table below presents a comprehensive overview of hyperparameter tuning results obtained from a grid search conducted on a sentiment analysis task. The configurations explore a combination of different learning rates, batch sizes, and epoch counts, along with the distinction between two model architectures: `roberta-simple` and `roberta-gru`. Notably, some configurations involve unfreezing layers during training, which can lead to significant differences in performance.

No.	Model Type	Learning Rate	Train BS	Epochs	F1 Score	Test Acc.
1	roberta-simple	1e-05	16	3	0.73	0.71
2	roberta-simple	2e-05	32	3	0.73	0.71
3	roberta-simple	1e-05	16	6	0.77	0.75
4	roberta-gru	1e-05	16	3	0.79	0.78
5	roberta-gru	2e-05	32	3	0.79	0.77
6	roberta-gru	1e-05	16	6	0.80	0.78
7	roberta-simple	3e-05	32	4	0.76	0.74
8	roberta-gru	3e-05	32	4	0.79	0.77
9	roberta-simple	1e-05	64	5	0.71	0.69
10	roberta-gru	1e-05	64	5	0.78	0.77
11	roberta-gru(6 Layer Unfreezing)	1e-05	32	2+8	0.86	0.85
12	roberta-gru (12 Layer Unfreezing)	1e-05	32	2+8	0.87	0.87
13	roberta-simple(6 Layer Unfreezing)	1e-05	32	2+8	0.85	0.84
14	roberta-gru (12 Layer Unfreezing)	1e-05	32	2+8	0.86	0.86

Table 1: Hyperparameter Tuning Results

From the results, it is evident that both model types benefit from a fine-tuned learning rate of $1e-05$, which consistently yields higher F1 scores across various configurations. The `roberta-gru` model, in particular, shows a marked improvement in performance with the increase in epochs and the unfreezing of layers, reaching an F1 score as high as 0.87406 and test accuracy peaking at 0.87.

The batch size seems to have a less consistent impact on performance, but larger batch sizes combined with a higher number of epochs and layer unfreezing tend to yield better results. For instance, configuration 12, employing the `roberta-gru` model with 12 layers unfrozen and an extended training period of $2+8$ epochs, achieves the highest F1 score and test accuracy.

Given the data, the optimal hyperparameters for this specific sentiment analysis task are as follows:

- **Model Type**: `roberta-gru`
- **Learning Rate**: $1e-05$
- **Training Batch Size**: 32
- **Epochs**: 10 (interpreted as $2+8$ from the provided configurations)
- **Unfreeze Layers**: 12

The choice to unfreeze 12 layers indicates that allowing more flexibility in the pre-trained model's parameters contributes significantly to the model's ability to learn from domain-specific data. However, it is crucial to validate these parameters on an independent test set to ensure the model has not overfitted to the validation set used during tuning.

In conclusion, the `roberta-gru` architecture, with carefully chosen hyperparameters and a strategy that includes unfreezing layers, offers a robust approach to improving sentiment analysis performance on the dataset.

6 Experiment and Result Analysis

This section presents a detailed analysis of the experiments conducted with the RoBERTa-GRU model under different training conditions, focusing on key metrics such as F1 score, test accuracy, and AUC scores for each sentiment class. The optimal hyper parameter values obtained after grid search were used to conduct these experiments. Also, the dataset now

6.1 With Dataset Augmentation and Balancing

In the first experiment, the dataset was augmented by downsampling the positive class to match the average size of the other classes. The results were as follows:

- F1 Score: 0.885
- Test Accuracy: 0.88
- AUC Scores: Negative - 0.985, Neutral - 0.897, Positive - 0.975

These results indicate a balanced and effective model performance across all classes, underscoring the benefits of dataset augmentation and balancing.

6.2 Without Dataset Augmentation and Balancing

The second experiment, conducted without dataset augmentation and balancing, yielded these results:

- F1 Score: 0.883
- Test Accuracy: 0.89
- AUC Scores: Negative - 0.986, Neutral - 0.902, Positive - 0.977

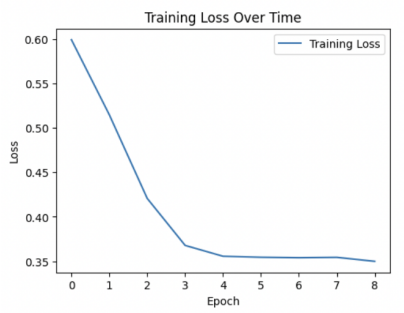
Despite the absence of data balancing, the model achieved a slightly higher test accuracy. However, the F1 score was marginally lower than in the balanced dataset experiment. A key factor in this performance is the imbalance in the test dataset, which mirrors the training dataset's skew towards the positive class.

6.3 Gradual Unfreezing of Pretrained RoBERTa Layers

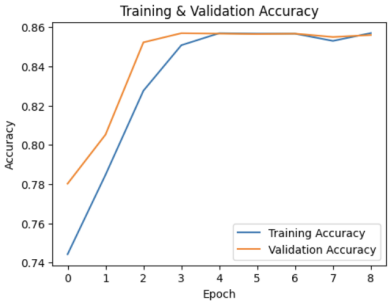
Another experiment involved gradually unfreezing the pretrained RoBERTa layers:

- F1 Score: 0.827
- Test Accuracy: 0.81
- AUC Scores: Negative - 0.967, Neutral - 0.835, Positive - 0.952

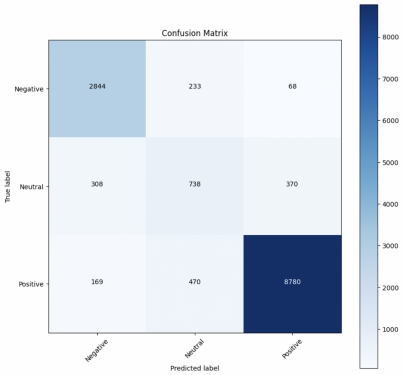
This strategy led to lower performance metrics, suggesting that gradual unfreezing did not favorably impact the model's accuracy for this specific task.



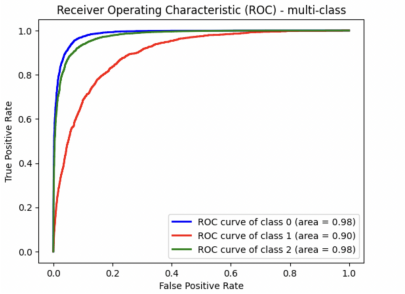
(a) Training Epoch vs Training Loss



(b) Training accuracy vs Test accuracy

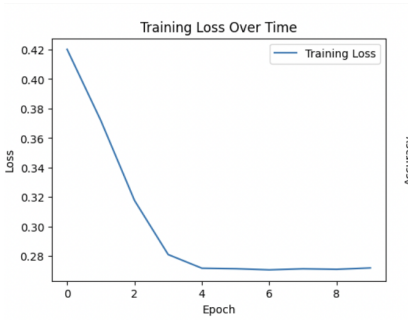


(c) Confusion Matrix

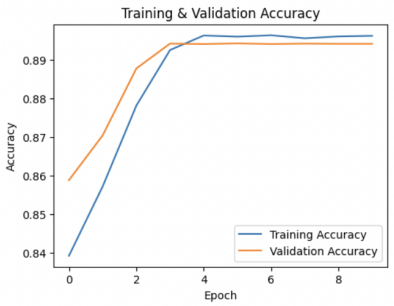


(d) AUC vs ROC Curve

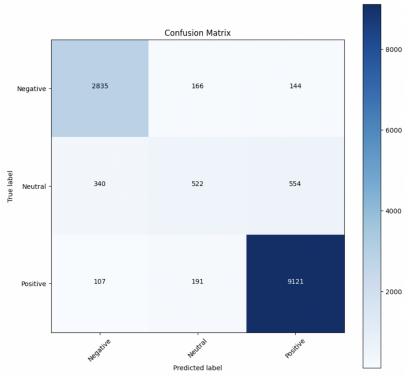
Figure 4: Test Set Performance Plot for Experiment 1



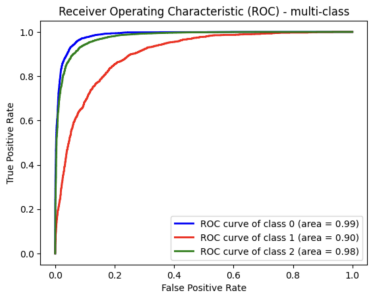
(a) Training Epoch vs Training Loss



(b) Training accuracy vs Test accuracy

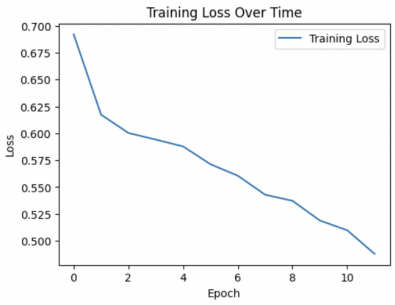


(c) Confusion Matrix

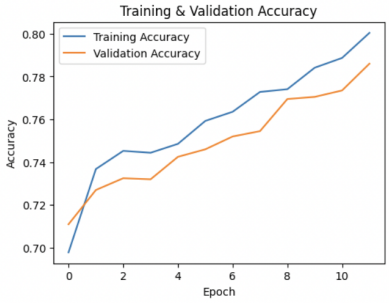


(d) AUC vs ROC Curve

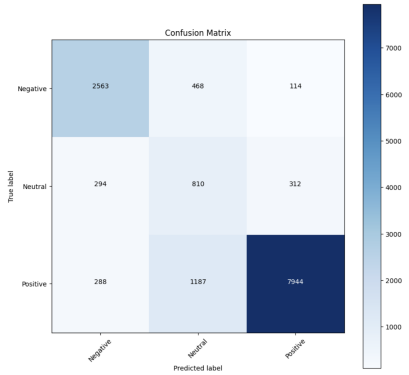
Figure 5: Test Set Performance Plot for Experiment 2



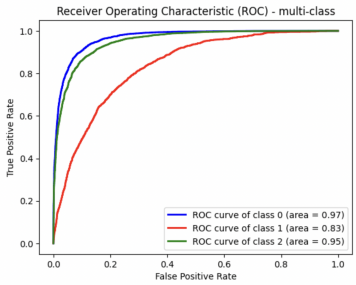
(a) Training Epoch vs Training Loss



(b) Training accuracy vs Test accuracy



(c) Confusion Matrix



(d) AUC vs ROC Curve

Figure 6: Test Set Performance Plot for Experiment 3

6.4 Analysis

These experiments reveal crucial insights into the RoBERTa-GRU model's behavior. Data balancing and augmentation improve the model's ability to generalize across different sentiment classes. In contrast, the model's favorable performance in the imbalanced dataset experiment can be attributed to the similar imbalance in the test dataset, particularly with an overrepresentation of positive reviews. Such findings highlight the necessity of aligning training and testing conditions for a fair assessment of model performance and emphasize the importance of dataset composition in sentiment analysis tasks.

7 Conclusion

This project successfully demonstrated the efficacy of a RoBERTa-GRU hybrid model in the sentiment analysis of Yelp restaurant reviews. Key takeaways include:

- The RoBERTa-GRU integration excelled in understanding contextual nuances and long-range dependencies in text, highlighting its effectiveness for complex NLP tasks.
- Data augmentation and balancing significantly improved the model's accuracy across diverse sentiment classes, emphasizing the importance of dataset composition in training.
- Optimal hyperparameters, identified through grid search, enhanced learning efficiency and overall model performance.
- A two-phase training strategy provided deeper insights into the adaptability and learning capabilities of the model.

The high F1 and AUC scores, particularly in the balanced dataset, attest to the model's robustness. However, the dependency of model performance on dataset characteristics suggests a potential area for further refinement. Future research could focus on advanced data preprocessing techniques, more nuanced data balancing approaches, and the integration of richer contextual features to elevate the model's analytical precision and reliability in sentiment analysis.

References

[1] T Hemakala and S Santhoshkumar. Advanced classification method of twitter data using sentiment analysis for airline service. In *Int. J. Comput. Sci. Eng.*, volume 6, pages 331–335, 2018.

[2] U Makhmudah, S Bukhori, JA Putra, and BAB Yudha. Sentiment analysis of indonesian homosexual tweets using support vector machine method. In *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, pages 183–186, Jember, Indonesia, 2019. IEEE.

[3] H AlSalman. An improved approach for sentiment analysis of arabic tweets in twitter social media. In *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, pages 1–4, Riyadh, Saudi Arabia, 2020. IEEE.

[4] A Gupta, A Singh, I Pandita, and H Parashar. Sentiment analysis of twitter posts using machine learning algorithms. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 980–983, New Delhi, India, 2019. IEEE.

[5] M Rhanoui, M Mikram, S Yousfi, and S Barzali. A cnn-bilstm model for document-level sentiment analysis. *Mach. Learn. Knowl. Extr.*, 1:832–847, 2019.

[6] V Tyagi, A Kumar, and S Das. Sentiment analysis on twitter data using deep learning approach. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 187–190, Greater Noida, India, 2020. IEEE.

[7] B Jang, M Kim, G Harerimana, Su Kang, and JW Kim. Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. *Appl. Sci.*, 10:5841, 2020.

[8] N Hossain, MR Bhuiyan, ZN Tumpa, and SA Hossain. Sentiment analysis of restaurant reviews using combined cnn-lstm. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5, Kharagpur, India, 2020. IEEE.

[9] P Harjule, A Gurjar, H Seth, and P Thakur. Text classification on twitter data. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pages 160–164, Jaipur, India, 2020. IEEE.

[10] Kian and Chin. Roberta-gru: A hybrid deep learning model for enhanced sentiment analysis. *Applied Sciences*, 13(6):3915, 2023.