Yelp Reviews Classification using Transformer

Faizan Ali Khaji

December 13, 2023

1 Abstract

This paper introduces a novel model for Yelp Review classification, leveraging a pretrained BERT model integrated with an LSTM layer and a subsequent Feed Forward Neural Network. The model was trained and evaluated on the Yelp reviews dataset, encompassing reviews accompanied by ratings ranging from one to five. The proposed model achieved an impressive accuracy of 89% on the test dataset. Additionally, the paper involves training model with varying numbers of attention heads and hidden layers, to understand the impact of these architectural parameters on model performance

2 Introduction

Sentiment analysis, a critical task in natural language processing (NLP) or opinion mining, involves extracting attitudes, evaluations, thoughts, and opinions expressed towards a specific subject, particularly in the context of the vast amount of user-generated unstructured text on the web, notably in social media [4]. Traditional machine learning models for sentiment analysis heavily rely on labeled datasets and shallow classification methods, often encountering challenges related to feature engineering, high dimensionality, and data sparsity, especially with the prevalent bag-of-words model [2].

Recurrent Neural Networks (RNNs) have gained popularity for NLP tasks due to their success with Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures[3]. However, these models face limitations, such as difficulties in parallelization and handling long clauses due to vanishing gradient issues. To address these challenges, the Transformer model, introduced with an encoder-decoder structure, offers a novel approach. The Transformer employs a Multi-Head Attention layer shared between the Encoder and Decoder, utilizing queries, keys, and values to generate weighted sums[6].

Furthermore, the Bidirectional Encoder Representations from Transformers (BERT) model, introduced by Google in 2018, has revolutionized NLP with its groundbreaking bidirectional attention mechanism. BERT's multi-layered transformer architecture enables it to excel in sentiment analysis and other

language understanding tasks by capturing intricate linguistic nuances through consideration of the entire context of words[1].

This paper proposes a methodology for yelp reviews classification by combining BERT's contextualized word representations with an LSTM. The objective is to classify news articles as positive, negative, or neutral, using accuracy, precision, recall, and F1 Score as evaluation criteria. Additionally, the study explores the impact of different configurations, such as varying hidden layers and attention heads, on the model's performance.

3 Methodology

The architecture proposed for this study entails stacking LSTM over the BERT architecture, which has demonstrated good results in fake news classification[5]. This architecture is adapted for Yelp review classification to perform multi-label sentiment analysis. Each review undergo preprocessing before passing through the BERT model to obtain contextual embeddings, which are subsequently processed through an LSTM layer for classification.

3.1 Dataset Overview

The Yelp reviews dataset comprises 174,000 reviews with star ratings in the training set and 13,980 reviews in the test set. The following plot illustrates the distribution of the training data.

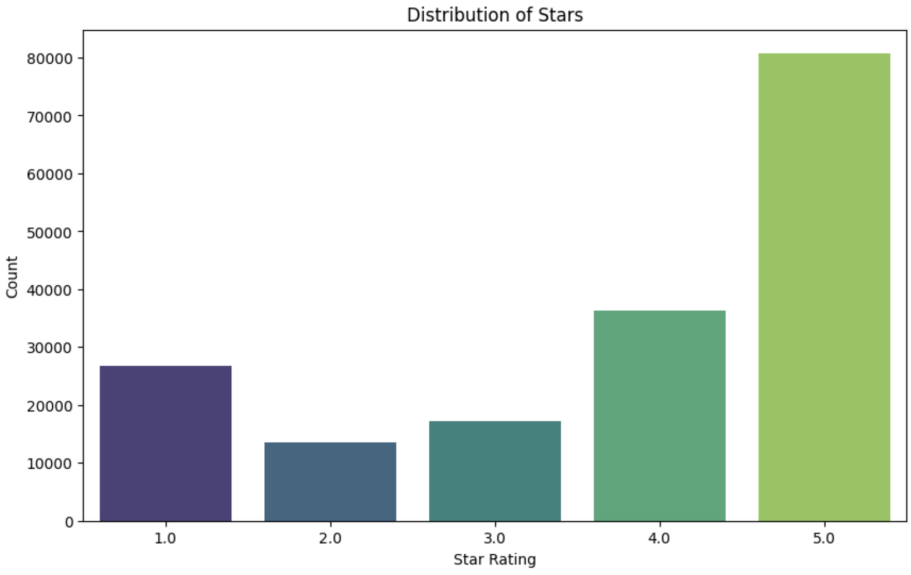

Figure 1: Distribution of Yelp Reviews training Dataset.

3.2 Data Preprocessing

Data preprocessing stands as a crucial phase in training models; hence, all reviews within the dataset undergo preprocessing. This involves the removal of stop words and punctuation, standardization to lowercase, and subsequent tokenization using the BERT tokenizer, followed by conversion to word encodings. Ensuring a consistent sequence length is essential for the BERT model. To achieve this standardization, reviews undergo either padding, in the case of a sequence length below the specified maximum, or truncation for sequences that exceed it.

To determine the optimal maximum sequence length for the BERT model, an analysis of the sequence length distribution in the dataset was done. Using box plot and histogram as shown in Figure 2 and Figure 3 aided in understanding the typical length and identifying outliers, guiding the establishment of a optimal maximum. Based on the plots as sequence length of 147 encompasses about 95% of the dataset, establishing it as the best sequence length for efficient training.

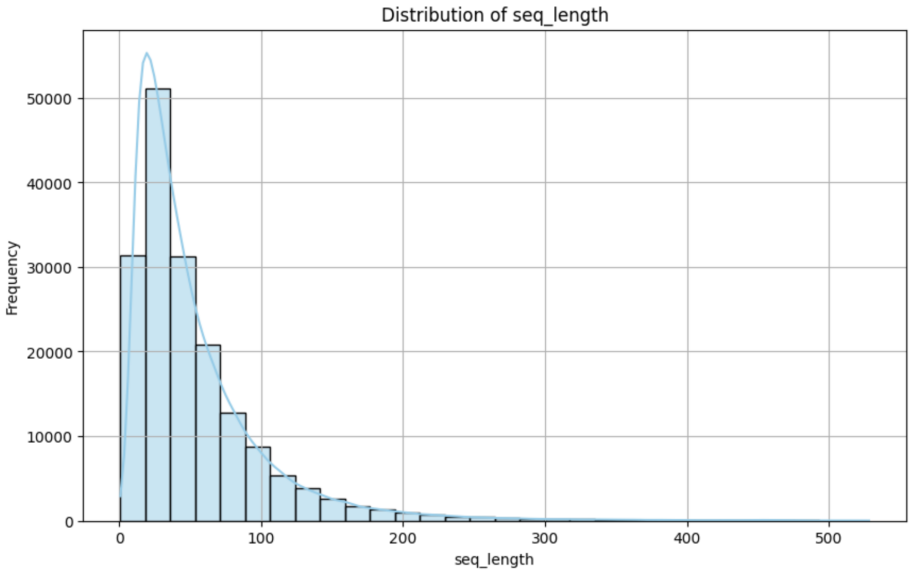

Figure 2: Histogram of sequence length.

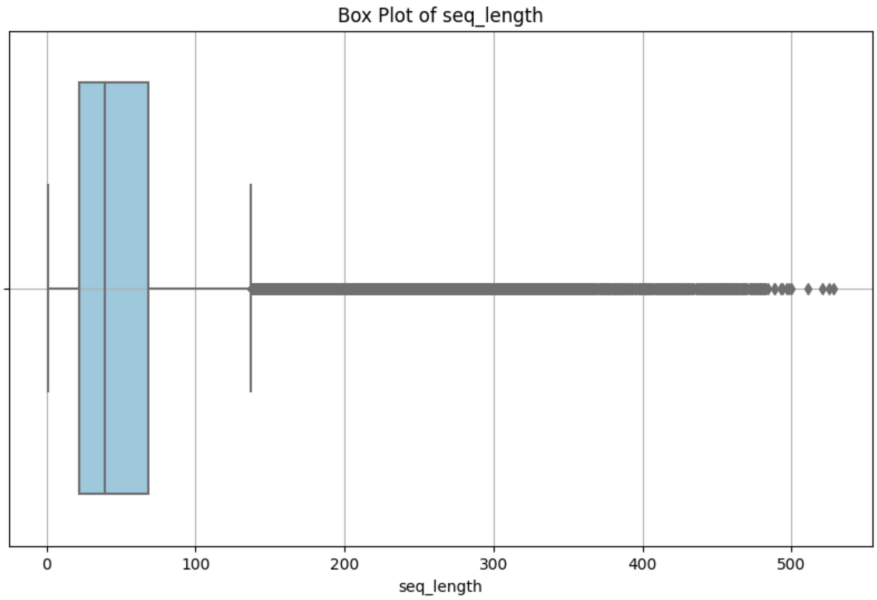

Figure 3: Box plot of sequence length.

3.3 Bert Model

BERT offers various pre-trained models tailored for diverse applications, with two widely used versions being $\text{BERT}_{\text{Base}}$ (L=12, H=768, A=12, Total Parameters=110M) and $\text{BERT}_{\text{Large}}$ (L=24, H=1024, A=16, Total Parameters=340M), each characterized by specific layer, hidden size, and attention head configurations. Within this spectrum, BERT-Base Uncased, a smaller variant (L=12, H=768, A=12, Total Parameters=110M), is employed in the proposed architecture. This model, pre-trained on uncased text, strikes a balance between computational efficiency and performance, serving as a foundational component for tasks like sentiment analysis and question answering. In the proposed architecture, BERT-Base-Uncased is utilized to generate high-dimensional vectors capturing the contextual information of each token within the input sequence, providing contextualized word embeddings for downstream tasks.

3.4 LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) designed to capture and process sequential information effectively. LSTMs address the vanishing gradient problem of traditional RNNs, enabling them to retain and utilize information over longer sequences. With memory cells and gates, LSTMs can selectively store and access information, making them well-suited for tasks involving sequential data.

In the proposed architecture, LSTMs are employed with 128 hidden units and one layer before passing the output to a feed-forward neural network. This configuration enhances the model's capacity to capture intricate relationships within sequential data.

When stacking LSTM with BERT, the LSTM layer aids in capturing contextual dependencies and long-range dependencies within the input sequence, complementing BERT's ability to understand context at the word level. This approach helps the model effectively classify and analyze sequential data, contributing to improved performance in natural language processing tasks, such as sentiment analysis or classification of textual data.

3.5 Model Architecture

In the proposed architecture, the BERT-base-uncased model and LSTM are leveraged alongside a feed-forward network. After preprocessing, Yelp reviews undergo tokenization, converting the tokens into numerical representations, and addressing sequence length requirements for the BERT model, resulting in input IDs and attention masks using the BERT tokenizer. These inputs are fed into the BERT model, generating 768-dimensional embeddings for each token. BERT provides contextualized sentence-level representations. Due to the bidirectional nature of BERT, the embeddings from the BERT model are input to LSTM, which consists of one layer with 128 hidden units. LSTM returns the output for each word in the sequence and is then fed into a feed-forward neural network for classification.

Two dense layers followed by the output layer helps in classifying the review. Notably, dropout layers are added over each layer to enhance generalization, and ReLU activation functions over each layer in the feed-forward neural network introduce non-linearity. This model design balances the power of transformer-based contextualization with sequential learning capabilities, presenting a versatile framework for various NLP applications.

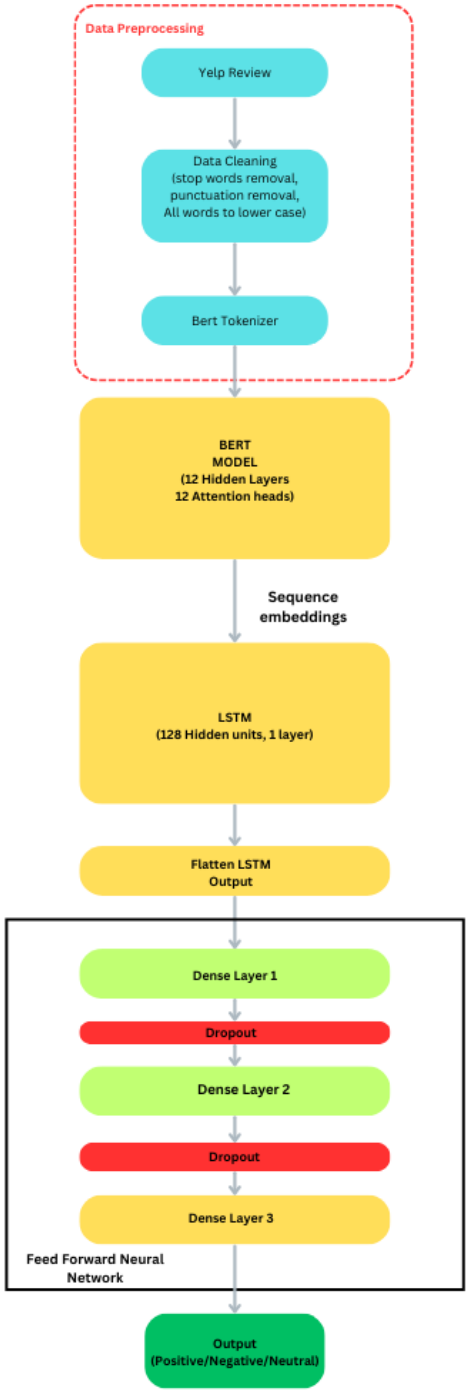

Figure 4: Model Architecture for yelp review classificaiton

4 Experiments and results

The model, combining BERT and LSTM, was fine tuned on the Yelp reviews dataset. To evaluate the model's performance on the test set, a confusion matrix was generated, as illustrated in Figure 5.The classification report offers a detailed analysis of the model's performance across different sentiment classes. Notably, the model achieves a commendable precision of 0.86 for the Negative class, indicating its accuracy in correctly identifying negative sentiments, and a high recall of 0.90 signifies its ability to capture the majority of actual negative instances. In contrast, the model's performance is less pronounced in the Neutral class, with a precision of 0.60 and recall of 0.39, suggesting challenges in accurately classifying neutral sentiments. However, the model excels in the Positive class, demonstrating outstanding precision (0.93), recall (0.97), and F1-score (0.95), showcasing its robust capability in identifying positive sentiments.

Classification Report:

	precision	recall	f1-score	support
Negative	0.86	0.90	0.88	3145
Neutral	0.60	0.39	0.47	1416
Positive	0.93	0.97	0.95	9419
accuracy			0.89	13980
macro avg	0.80	0.75	0.77	13980
weighted avg	0.88	0.89	0.88	13980

Figure 5: Classification matrix

The overall accuracy of the model stands at 89%, reflecting its effectiveness in classifying sentiments across the entire dataset. The macro and weighted averages provide a comprehensive view of the model's performance, highlighting its proficiency in distinguishing between positive and negative sentiments while revealing areas for potential improvement in accurately classifying neutral sentiments Additionally, a visualization depicting the precision of the model on the test set has been created.

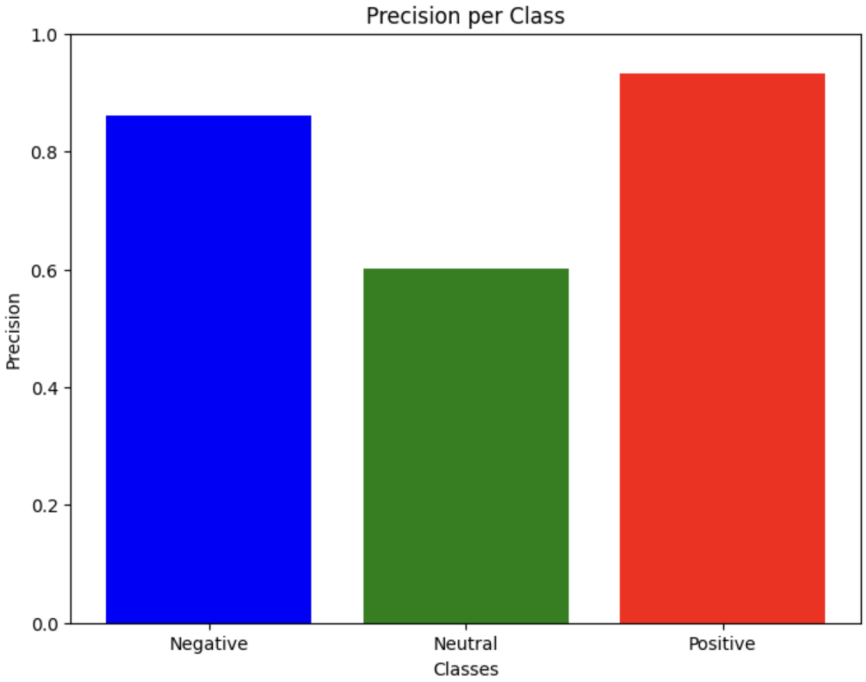

Figure 6: Precision per class

4.1 Experiment - Impact of Number of Attention heads and Number of hidden layers in BERT

Throughout the epochs, the model exhibited a rapid tendency to overfit the training data. To enhance generalization, a few layers and attention heads from the BERT model were dropped. While this adjustment did not result in a substantial improvement in model accuracy on the validation data, it significantly reduced the time required for fine-tuning. This optimization facilitates faster model training without significantly compromising its performance during testing. The below images illustrate the effects of these modifications on accuracy, loss, and the total time taken by the model.

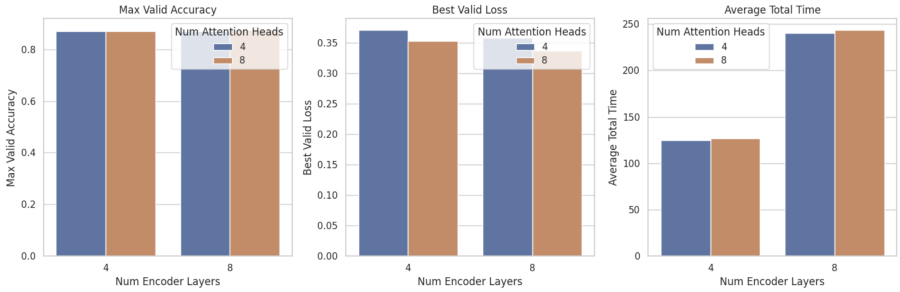

Figure 7: Impact on model with different attention heads and hidden layers

5 Conclusion

In conclusion, the model, integrating BERT and LSTM, presents a robust framework for sentiment analysis. The optimized architecture successfully balances the power of transformer-based contextualization with sequential learning capabilities. Achieving an accuracy of 89%, the model demonstrates proficiency in classifying sentiments within Yelp reviews.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] A. I. Kadhim. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52:273–292, 2019.

[3] Aytuğ Onan. Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *Journal of King Saud University - Computer and Information Sciences*, 34(5):2098–2117, 2022.

[4] Aytuğ Onan. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, 33(23):e5909, 2021. e5909 CPE-20-0130.R1.

[5] Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. Fake news classification using transformer based enhanced lstm and bert. *International Journal of Cognitive Computing in Engineering*, 3:98–105, 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you

need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.