Using Deep Learning to Identify Sentiment in IMDb Movie Reviews

JIAHAO LIU
UCD School of Computer Science
University College Dublin Belfield, Dublin 4, Ireland.
jiahao.liu@ucdconncet.ie

Abstract

Sentiment analysis is a critical task in natural language processing that involves determining the sentiment expressed in a given text, such as positive or negative. In this project, I propose a deep learning model, Bi-GRUConv, that combines bidirectional Gated Recurrent Units (Bi-GRUs) and Convolutional Neural Networks (CNNs) for sentiment analysis in movie reviews. The model aims to capture both the long-range dependencies and hierarchical structure in text data, and employs pretrained Word2Vec embeddings for improved performance. The experimental setup includes a thorough preprocessing of the IMDb movie reviews dataset, followed by model building and training, incorporating various optimization techniques such as early stopping and learning rate reduction to prevent overfitting. I compare the performance of the Bi-GRUConv model with a traditional LSTM-based model, evaluating both models on a test set and reporting accuracy and F1 score metrics to provide a comprehensive assessment of their effectiveness in sentiment classification. The results demonstrate that the Bi-GRUConv model outperforms the LSTM model, showing the benefit of combining Bi-GRUs and CNNs for sentiment analysis tasks.

1 Introduction

Sentiment analysis is a crucial task in natural language processing that involves identifying the emotions or feelings conveyed in a given text, such as reviews of movies or products (Rohman et al., 2020). Precise sentiment analysis can offer valuable insights into customer preferences, enabling businesses to make informed, data-driven decisions (Ardakani et al., 2021). The emergence of deep learning techniques has facilitated the development of more efficient sentiment analysis models by utilizing their ability to recognize intricate patterns from raw data (Araque et al., 2017).

In this project, I introduce a deep learning model that merges bidirectional Gated Recurrent Units (Bi-GRUs) and Convolutional Neural Networks (CNNs) to differentiate movie reviews as either positive or negative based on their written content. The proposed model capitalizes on the advantages of both Bi-GRUs and CNNs to capture the extensive dependencies and hierarchical structure in textual data. The effectiveness of this model is evaluated using the IMDb movie reviews dataset and compared to a conventional LSTM-based model.

The project is organized into different sections. In Section 2, there is a brief overview of the related work in sentiment analysis, which includes the use of deep learning models, bidirectional RNNs, and the combination of CNNs and RNNs. Section 3 explains the experimental setup, which includes the selection of datasets, preprocessing, model building and training, and evaluation. In Section 4, the results are presented, which includes evaluating the model's performance by testing it and calculating

both the test accuracy and F1 score. Lastly, Section 5 concludes and discusses potential future work in the field of sentiment analysis using deep learning techniques.

2 Related Work

2.1 Sentiment Analysis with Deep Learning Models

Sentiment analysis is a well-researched topic in natural language processing, and deep learning models have been incredibly successful in this area. Several models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Unit (GRU) networks, have demonstrated promising outcomes in sentiment analysis tasks. For example, LSTM networks have been effectively utilized for sentiment classification in movie reviews (Jang et al., 2020), and have reached state-of-the-art performance.

2.2 Bidirectional Gated Recurrent Unit (Bi-GRU) and LSTM Models

Bidirectional Gated Recurrent Unit (Bi-GRU) models and Bidirectional LSTM models, variants of RNNs, have also been employed for sentiment analysis tasks. Both models are particularly useful in capturing forward and backward dependencies in the input sequences (Chung et al., 2014). In recent years, researchers have demonstrated the effectiveness of Bi-GRU and LSTM models in various sentiment analysis tasks, such as movie review classification social media sentiment analysis, and opinion mining in product reviews (Severyn and Moschitti, 2015).

2.3 Combining CNNs and RNNs for Sentiment Analysis

Combining the strengths of CNNs and RNNs, such as GRUs and LSTMs, has been explored in the literature for sentiment analysis. For example, Wang et al. (2016) proposed a CNN-RNN model that integrates the hierarchical structure of text data using CNNs and captures long-range dependencies using RNNs. This combination has been shown to improve the performance of sentiment classification tasks in multiple domains, such as French customer reviews in Amazon (Habbat et al., 2023).

2.4 Transfer Learning and Pretrained Word Embeddings

In the field of sentiment analysis, transfer learning has been widely adopted to enhance model performance. Pretrained word embeddings, in particular, have proven to be effective. Word2Vec (Mikolov et al., 2013) is a commonly used method for generating pretrained word embeddings that can be fine-tuned for specific tasks. These embeddings have been utilized in various deep learning models, including CNNs and bidirectional RNNs like GRUs and LSTMs, to improve their performance in sentiment analysis.

3 Experimental Settings

3.1 dataset

This project utilizes the IMDb movie reviews dataset from Kaggle, which includes 50,000 movie reviews that are equally divided into positive and negative sentiment labels. Each review in the dataset is labeled with either 'positive' or 'negative' sentiment, indicating favorable or unfavorable reviews, respectively. Figure 1 and 2 show the distribution of comments based on the number of words and comments in each corresponding interval. Additionally, Figure 3 and 4 displays a word cloud of the dataset.

3.2 Preprocessing

Text preprocessing is an essential step in making texts more manageable for natural language processing, particularly for classification tasks. This process entails removing stop words, special

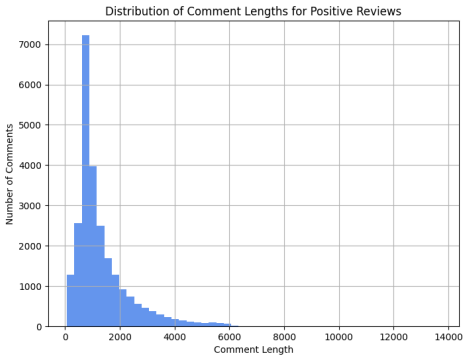

Figure 1: The distribution of positive comments

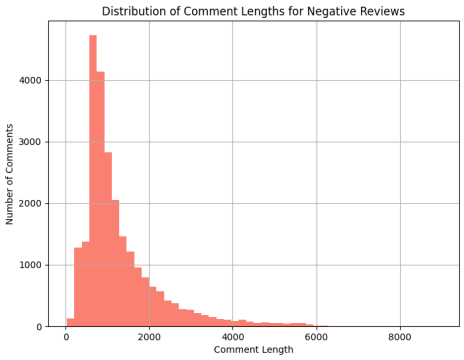

Figure 2: The distribution of negative comments

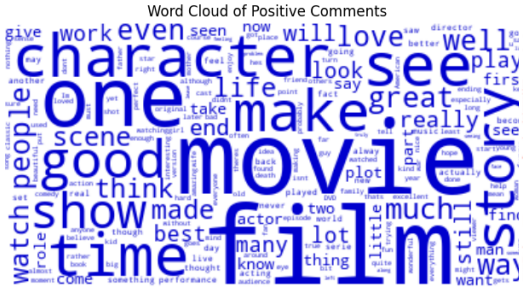

Figure 3: The positive comments word cloud

Figure 4: The negative comments word cloud

characters, and HTML tags, as well as standardizing words. By doing so, the input data becomes more organized and relevant to ensure better and more accurate classification results (Kasliwal et al., 2018).

To prepare the dataset for training and modeling, various preprocessing steps were taken. Firstly, the review text was cleaned, removing any HTML tags, special characters, URLs, email addresses, numbers, and extra white spaces. The text was tokenized next, with all of the tokens converted to lowercase. These tokenized reviews were then used to create word embeddings in a Word2Vec model. The reviews were then transformed into sequences using the Keras tokenizer, with a maximum of 10,000 words. These sequences were then padded to ensure a consistent length of 500 tokens.

3.3 Model Building Training

The Conv1D model is better suited for sequential data such as text. This is because it can easily capture local patterns in the data. To perform binary sentiment classification, a deep learning model is used which combines bidirectional GRU and CNN layers. The model starts with an Embedding layer that uses pretrained Word2Vec embeddings, followed by a Dropout layer. Next, a Bidirectional GRU layer with 128 units is added, followed by a Batch Normalization layer. To enhance the model's performance, a 1D CNN layer is added with 128 filters and a kernel size of 3 , which is succeeded by a GlobalMaxPooling1D layer. Finally, the binary output is generated by the Dense layer, which has a sigmoid activation function, which can be seen in Figure 5.

I added another model that uses an LSTM layer in Figure 6 to compare its performance with the previous architecture . LSTM layers are suitable for text classification tasks as they can manage long-term dependencies in sequential data . The comparison between bidirectional GRU and LSTM models will help us better comprehend the efficacy of each architecture in addressing the sentiment classification issue.

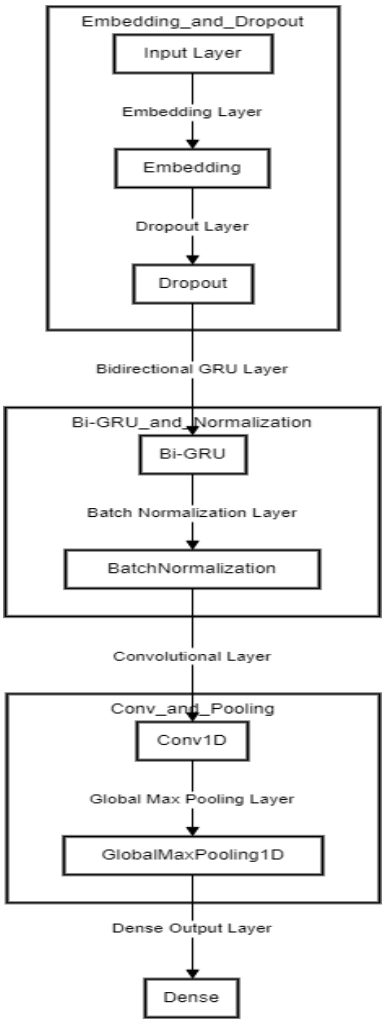

Figure 5: The Bi-GRU-CONV model

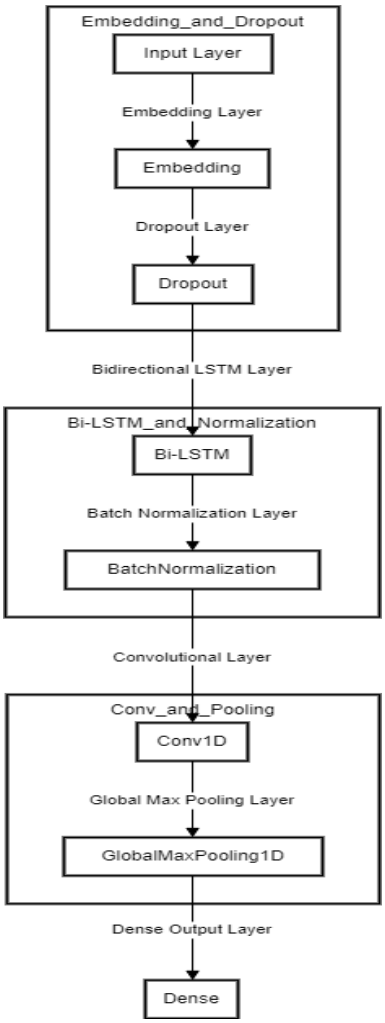

Figure 6: The LSTM model

I selected the Adam algorithm for optimization as it can efficiently manage sparse gradients and noisy problems. To compile the model, I used the binary_crossentropy loss function and accuracy metric. During the training process, I utilized a batch size of 64 and a maximum of 30 epochs, implementing early stopping and learning rate reduction to prevent overfitting and enhance the model's performance. The dataset was divided into 80% for training and 20% for testing, with 10% of the training data serving as validation during the training process.

3.4 Evaluation

Last, I evaluated the model's performance using the test set and calculated both the test accuracy and F1 score to actively assess the classification performance.

This section presents the results of two deep learning models, namely Bidirectional Gated Recurrent Unit (Bi-GRU) and Long Short-Term Memory (LSTM) networks. The results consist of the final test accuracy and F1 score for each model.

The Bi-GRU model achieved a test accuracy of 0.9208 and an F1 score of 0.9218 after 29 epochs of training. The model's training process included adjusting the learning rate using the ReduceLROnPlateau callback, which helped improve its performance. The best validation accuracy achieved during training was 0.9175 at epoch 23, which can be seen in Figure 7 and 8.

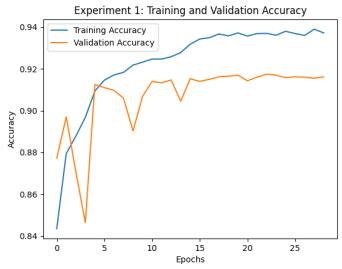

Figure 7: Training and Validation Accuracy 1

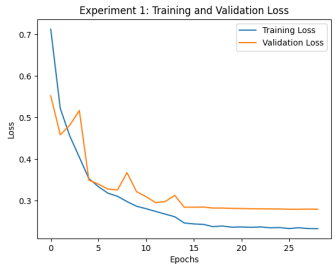

Figure 8: Training and Validation Loss 1

The LSTM model underwent 27 epochs of training and was able to attain a test accuracy of 0.9171 and an F1 score of 0.9183. Just like the Bi-GRU model, the learning rate was regulated via the ReduceLROnPlateau callback. The best validation accuracy noted during training was 0.9202 at epoch 12, which can be seen in Figure 9 and 10.

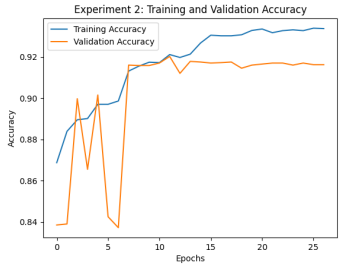

Figure 9: Training and Validation Accuracy 2

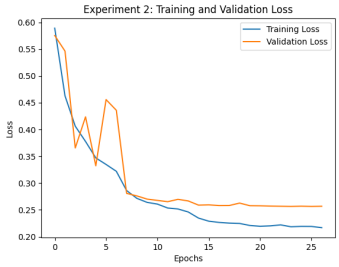

Figure 10: Training and Validation Loss 2

When I compare the two models, the Bi-GRU model was slightly more successful in achieving a higher test accuracy of 0.9208 and an F1 score of 0.9219 as opposed to the LSTM model's 0.9171 and 0.9183, respectively in Table 1. This suggests that the Bi-GRU model is better suited for this task, although other factors such as computational resources and training time must also be taken into account when selecting a model to deploy in a production environment.

Model	Test Accuracy	F1 score
Bidirectional GRU	0.9208	0.9219
LSTM	0.9171	0.9183

Table 1: Table:Accuracy and F1 score

4 Conclusion and Future Work

In conclusion, my project comparing the Bi-GRUConv and LSTM models for text classification shows that the Bi-GRUConv model outperforms the LSTM model in terms of test accuracy and F1 score. The Bi-GRUConv model's combination of bidirectional GRU layers and Conv1D layers may be the reason behind its ability to comprehend more complex patterns in the data.

For future work, we can analyze the performance of Bi-GRU, Conv, and LSTM models on different datasets and tasks. We can explore architectural enhancements and evaluating their scalability, which can contribute to developing more robust and effective text classification models in natural language processing (Zhang et al., 2015).

References

Rohman, A.N., Luviana Musyarofah, R., Utami, E. Raharjo, S. 2020, "Natural Language Processing on Marketplace Product Review Sentiment Analysis", IEEE, , pp. 1.

Ardakani, S.P., Zhou, C., Wu, X., Ma, Y. Che, J. 2021, "A Data-driven Affective Text Classification Analysis", IEEE, , pp. 199.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J.F. Iglesias, C.A. 2017, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications", Expert systems with applications, vol. 77, pp. 236-246.

Jang, B., Kim, M., Harerimana, G., Kang, S. Kim, J.W. 2020, "Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism", Applied sciences, vol. 10, no. 17, pp. 5841.

Chung, J., Gulcehre, C., Cho, K. Bengio, Y. 2014, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", .

Severyn, A. Moschitti, A. 2015, "Twitter Sentiment Analysis with Deep Convolutional Neural Networks", ACM, , pp. 959.

Wang, W., Chen, L., Thirunarayan, K. Sheth, A.P. 2012, "Harnessing Twitter "Big Data" for Automatic Emotion Identification", IEEE, , pp. 587.

Habbat, N., Anoun, H. Hassouni, L. 2023, "Combination of GRU and CNN deep learning models for sentiment analysis on French customer reviews using XLNet model", IEEE engineering management review, vol. 51, no. 1, pp. 1-9.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. Dean, J. 2013, "Distributed Representations of Words and Phrases and their Compositionality", .

Kasliwal, N. 2018, Natural language processing with Python quick start guide: going from a Python developer to an effective natural language processing engineer, Packt Publishing, Birmingham.

Zhang, X., Zhao, J. LeCun, Y. 2015, "Character-level Convolutional Networks for Text Classification", .