Sentiment Analysis of Movie Reviews using LSTM

Mehul Deorao Ganjude **Zulfiya Amin saiyed** **Sai Rishith Reddy Gade**

The University of Texas at Arlington

mdg0551@mavs.uta.edu zxs9057@mavs.uta.edu Sxg967@mavs.uta.edu

Abstract

In recent years, deep learning has achieved great success in many fields, such as Computer Vision and Natural language processing (NLP). Compared to traditional machine learning methods, deep learning has a strong learning ability and can make better use of datasets for feature extraction. Because of its practicality, deep learning is becoming more and more popular. This paper mainly focusses on some deep learning algorithms like LSTM, LSTM with CNN, GRU and GloVe encoding on 50,000 reviews from IMDB dataset. LSTM (Long short-term memory) and GRU (Gated recurrent unit) have gates as an internal mechanism, which control what information to keep and what information to throw out. The CNN LSTM architecture uses CNN) layer for feature extraction on input data combined with LSTM to support sequence prediction. GloVe encoding derives the relationship between the words from statistics. To implement GloVe encoding the train and test accuracy obtained by LSTM, LSTM with CNN and GRU has been compared in this paper. The research is based on the IMDB dataset, which includes movie reviews and the labels that go with them (positive or negative) with the help of glove encoding method. The objective is to find the most accurate and generalized model possible.

Introduction

Deep learning was developed from Artificial Neural Network, and now it is a prevalent filed of Machine Learning. This paper shows how sentimental analysis is performed on a large dataset which contains 50,000 reviews from an IMDB movie review dataset. Movie reviews may provide an in-depth analysis of the movie which helps the viewers understand and know the complete picture of how the movie would be. When the data in reviews becomes larger, it is very crucial to automate the process to save time.

Sentimental analysis also called as opinion mining is an NLP (natural language processing) technique used to determine whether the data is positive, negative, or neutral. This technique is helpful to figure out how the viewers feel about a movie or any other issue. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback and understand customer needs. Due to the complexity of the human language, sometimes it becomes challenging to interpret the review.

The major two drawbacks are that keywords with several meanings depending on context might cause ambiguity, and the inability to categorize sentences that do not have unambiguous emotional keywords may infer that the phrase does not include any emotions. As a result, because it will be used in highly sensitive applications, the designed system should accommodate for these flaws and ensure reliable data classification. This research performs various neural networks on database of the IMDB movie review datasets to perform sentiment analysis and get an accurate output. Sentiment analysis is a type of machine learning that focuses on extracting subject information form reviews in the form of text.

Dataset Description

This is one of the most extensive IMDb movie review datasets. There are 50,000 movie reviews in the dataset, divided into two categories: positive and negative out of which 25,000 are test sets and 25,000 are train sets. We got this dataset via Keras, which encodes each review as a series of word indexes. A negative review has a score of <=4 out of 10, and positive review has a score of >=7 out of 10. No more than 30 reviews are included per movie.
The downloaded link contains the following folders,

- Test folder
- Train folder
- README.txt

For the glove encoding we have merged all the dataset combined and kept in one file along with their respective sentiments to analysis the sentiments for the random input. Along with this file when data is trained it will calculate score of each sentiment and will store on one file(glove.6B.50d) to predict the output.

Project Description

1. Description

The fast growth of internet and social media usage has altered decision-making, this is inseparable from the availa-

-bility of the most important source of information, user opinion. One source of information is information about a movie. The movie is a work of fiction including drama, action of stories, and contents and characters. But not all types of movies have the same quality. Therefore, before deciding to go for the movie you should first find out information about the movie based on reviews given by others on social media. The reviews can help to find out whether the movie has a quality worth going or not. Reading the entire review can take a long time, but if only a few reviews are read, the evaluation will be biased. To understand the opinions of the review, algorithms and programs are needed to process information and opinion data and to analyze the opinions of social media users called sentiment analysis. Sentiment analysis or opinion mining is a field of study that analyzes one's opinions, sentiments, evaluations, attitudes, and emotions from written language. Sentiment analysis is conducted to assess the review of an object whether it tends to the positive opinion or negative opinion.

There are many studies that have applied sentiment analysis to a review, even many papers and methodologies held competitions to identify the best methods for sentiment classification. Machine learning methods such as Naïve Bayes, Maximum Entropy (ME), and Support Vector Machine (SVM) are often used in finding models and features that are appropriate to the target problem. SVM and ME are complex models so training time is longer, while Naïve Bayes is a simple and fast model. But machine learning also has problems in extracting complex features and finding better types of features. Semantic features can reveal deep and implicit semantic relationships between words which can be more useful in the classification of sentiments. In this LSTM, LSTM with CNN, GRU and Glove encoding is used in the feature extraction process. The LSTM method is used in research in conducting sentiment analysis that has better results compared to conventional methods. This shows the LSTM method is suitable to be applied in sentiment analysis. For these reasons, the Long Short-Term Memory (LSTM) method will be applied for sentiment analysis in the review of movie. In this study, the LSTM method will be compared with the LSTM with CNN method. Furthermore, when we compared the accuracy between LSTM and LSTM with CNN, we saw results are showing some improvement in addition on CNN layer in LSTM. Then again CNN method compared with another method called GRU to check the best suitable method.

This technique is helpful to figure out how the viewers feel about a movie or any other issue. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback and understand customer needs.

2. Main references used for project

For our project we took the main reference from Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory by (Saeed Mian Qaisar 2020). In this paper they have implemented only LSTM method on IMDB dataset. They have analyzed the review using the LSTM classifier. Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews another paper we have reviewed by (Kamil Topal and Gultekin Ozsoyoglu 2016). In this paper, they have implemented HourGlass of Emotions Model for identify the emotions and to visualize the aggregated per-dimension emotion score of a movie's reviews, they use the movies emotion map. The third reference we have use Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews by (Md. Rakibul Haque and Salma Akter Lima 2019) they have used LSTM and CNN separately and compare the accuracy between these two methods. The last reference that we have used for our project is Entire Information Attentive GRU for Text Representation by (Guoxiu He, Wei Lu 2018) They present a new neural network structure that consists of attention mechanism applied to GRU model. When we refer to many papers related to this topic, we came across many methodologies like naïve Bayes, SVM by (Brandon Joyce, Jing 2019) and many but they couldn't achieve more accuracy as implementation of this methods on huge dataset is time and CPU consuming.

3. Difference in APPROACH/METHOD between your project and the main projects of your references

We used the nltk package to pre-process the data in our method. After that, three distinct deep learning models were utilized.

3.1 Pre-processing

The nltk and re libraries are used to pre-process the text content. The following actions were conducted during pre-processing. To begin, all the characters in the text were converted to lower case. Second, all the tabs were removed and replaced with blank spaces. In the third phase, all punctuation, special characters, and other characters were deleted, simply leaving the alphabets. The text was then transformed to tokens, and the tokenizer was used to eliminate the stop words. The stop words were obtained using the nltk library. The we applied the lemmatization process where we take individual tokens from a sentence, and we try to reduce them to their base form. When lemmatization and token generation has been done, we load the train and test dataset with all preprocess data.

3.2 Deep Learning Models

3.2.1) LSTM (LONG- SHORT TERM MEMORY)

Long short-term memory is an artificial recurrent neural network architecture used in field of deep learning. Unlike standard feedforward neural network, LSTM is applications to tasks such as unsegmented, connected handwriting recognition, speech recognition and IDSs. LSTM unit consists of cells, input gate, output gate and forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. It can remember the previously predicted possibility. LSTM has controlling knobs which controls the flow and mixing of input which brings more flexibility in controlling the output and giving us better results.

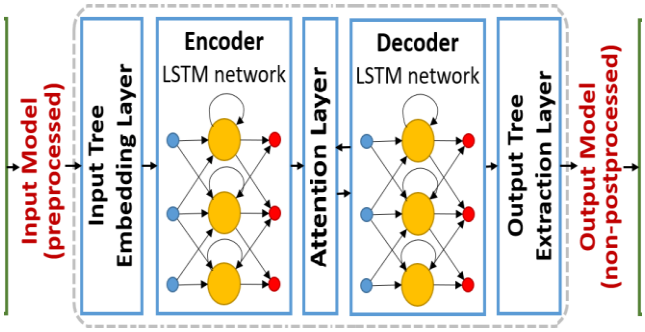

Figure 3.2.1.1

During training, the three gates figure out what information is important to keep and what information to forget. To build the LSTM model, first we have imported all the necessary libraries and we have used three layers. The embedding layer is the initial layer, and it passes the top words=4000, embedding vector length=32, and maximum review length=600 to the next layer as input. The LSTM layer, which has nodes=100 (It will take top 100 neurons), is the second layer. The Dense layer, which has only one node, where the sigmoid (value either 1 or 0) is used as the activation function, receives one output. The sigmoid function will provide a value between 0 and 1, allowing us to decide if the review is positive or negative. Along with this we have used the binary crossentropy as loss function for model compilation it will make sure that output will be in the range of 0 and 1. Adam optimizer is a very common optimizer for deep learning model therefore we have also chosen for our all model. This model is then trained for Epoch=8. By using LSTM for the IMDB movie review dataset and keeping epoch value=8 we get training accuracy as 93.92% and testing accuracy as 86.70.
The accuracy graph using the LSTM model is shown in Figure 3.2.1.2

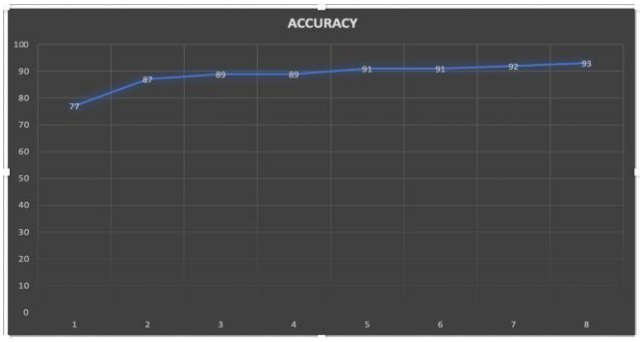

Figure 3.2.1.2

3.2.2) LSTM with CNN

A convolutional neural network is one that uses the convolutional process to figure out the relationships between two functions.

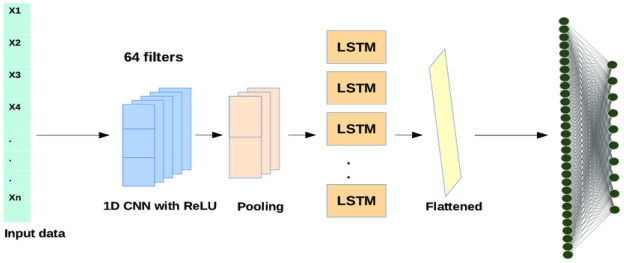

Figure 3.2.2.1

A 1D CNN layer is utilized as a pre-processing step to an LSTM layer, and the model is generated by mixing 1D CNN and LSTM. In particular, the 1D CNN layer is followed by a max-pooling layer with a window size of 3. After that, fully connected layers, such as a dense layer with sigmoid activation function are stacked. Adding CNN layer on the top of LSTM will give more accurate result because CNNs have the capacity to apply filters to compute dilations between each cell.

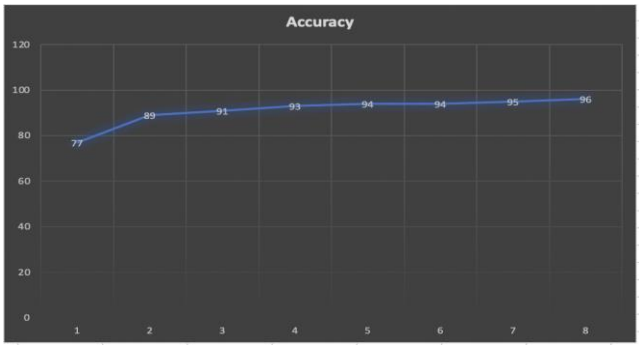

Figure 3.2.2.2

By adding CNN layer with LSTM for the IMDB movie review dataset and keeping epoch value=8 we get training accuracy as 97.16% and testing accuracy as 87.52. Accuracy has been increased by adding CNN layer.

The accuracy graph using the LSTM with CNN model is shown in Figure 3.2.2.2 The difference we observe when we add CNN layer between input and LSTM accuracy increases by 2-3% approx.

3.2.3) GRU (Grated Recurrent Unit)

GRU is like a long-short term memory with a forget gate but has fewer parameters than LSTM. It does not have an output gate but has update gate and reset gate. The main motive behind comparing above two model with GRU is that, to check the impact on accuracy when one gate is removed. GRU is a memory centered type of neural network other Neural Network does not have the ability to retain information. GRU is used when you have less memory consumption and wants faster result.

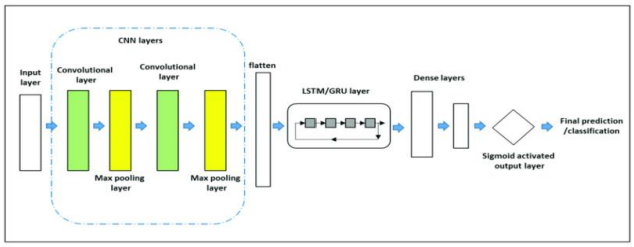

Figure 3.2.3.1

Before compiling the model GRU will Truncate and pad Input sequences to ensure that all sequences in a list have the same length. pad_sequences will take the maximum length and will compare with all length, if any length is shot of maximum length define then it will pad smaller length with 0. We have defined the maximum length as 500 and use to activation function relu and sigmoid for creating the model. For compiling the model, we have used same functions as we defined in above two models.

By using GRU for the IMDB movie review dataset and keeping epoch value=8 we get training accuracy=94.10% and testing accuracy=85.42.

There is not much difference between accuracy when GRU method compare with LSTM. But when it compares with LSTM with CNN there is some difference in the accuracy, and by this we can conclude that LSTM with CNN method is best among all the methods implemented on IMBD dataset.

The accuracy graph using the GRU model is shown in Figure 3.2.3.2

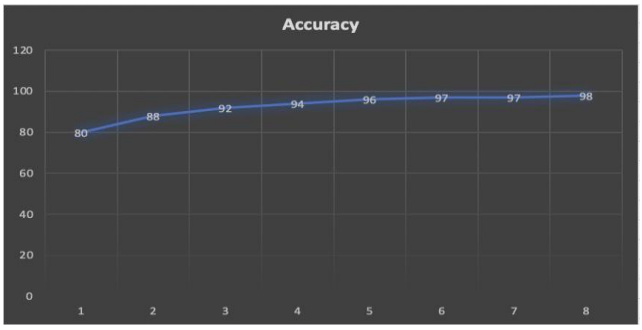

Figure 3.2.3.2

3.2.4) GloVe Encoding

GloVe stands for global vectors for word representation. Glove is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. The basic idea behind the GloVe word embedding is to derive the relationship between the words from statistics. Unlike the occurrence matrix, the co-occurrence matrix tells you how often a particular word pair occurs together. Each value in the co-occurrence matrix represents a pair of words occurring together.

For implementing Glove on IMDB dataset we have first combined all the data with their respective original sentiments. Moving ahead we have removed the all the stop words because Stop words are available in abundance in any human language. By removing these words, we remove the low-level information from our text to give more focus to the important information. When we removed the stop words, we compare our new data with old data and check whether all sentiment is correct. The model will calculate the score and if the prediction of score is > 0.5 then predicates sentiments is positive else is negative and if test score is = 1 then correct sentiment is positive else it is negative.

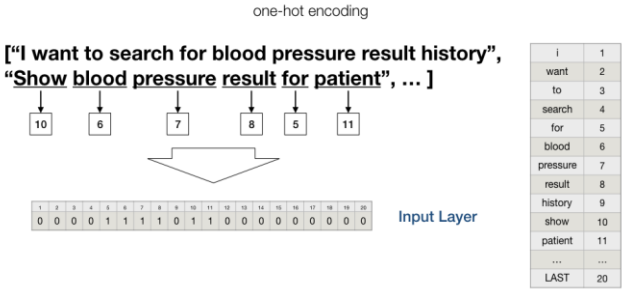

Figure 3.2.4.1

Figure 4.1 represents how the words has been dividend and kept in vector.

When we compile the model and kept the epoch value=8 we get training accuracy as 81.01% and testing accuracy as 84.15.

The paper what I refereed they have used only LSTM method on IMDB dataset and shows how LSTM can help to identify the sentiments and give the result. Long Short-Term Memory classifier is used with Adam optimizer to automatically categorize the preprocessed IMDb movie reviews. In total they have consider 10k reviews, 5k for positive and 5k for negative sentiments. But we have used the complete dataset and implement the model.

The second paper that I have referred introduced the notion of movie emotion maps for capturing the emotion content of movies via the emotions expressed in movie reviews. They have analyzed the characteristic of a set of movie reviews in the IMDb database, towards the goal of providing movie recommender systems based on the emotion maps of movies and the desired emotion maps of moviegoers.

The third paper that I have referred they implement the two models like us LSTM and LSTM with CNN on IMDB dataset, but they have implemented both the methods separately and compared the accuracy between them and shows which model will perform better on IMDB dataset. Furthermore, they have proposed for future work, they have decided to use the convolutional neural network in other fields of natural language processing and evaluate the performance of used methods in those fields.

4. Difference in ACCURACY/PERFOMANCE between your project and the main projects of your references

When we implemented the 3 models i.e., LSTM, LSTM with CNN and GRU, we saw there was increase in the accuracy when CNN layer implemented with LSTM but there is no more change between GRU when it compared with LSTM.

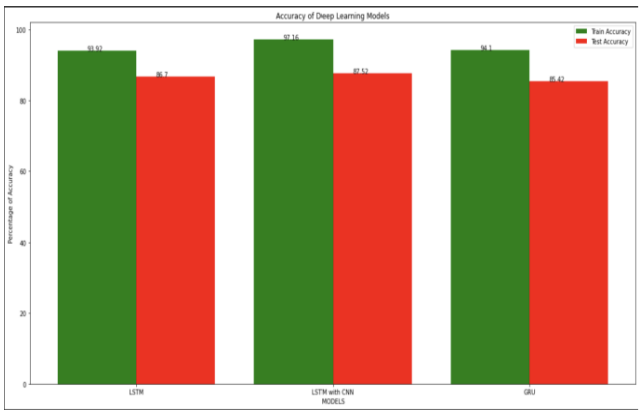

Figure 4

Gated Recurrent Units, in simple words, the GRU unit does not have to use a memory unit to control the flow of information like the LSTM unit. It can directly make use of the all-hidden states without any control. GRUs have fewer parameters and thus may train a bit faster or need less data to generalize. But, with large data, the LSTMs with higher expressiveness may lead to better results. They are almost like LSTMs except that they have two gates: reset gate and update gate. Reset gate determines how to combine new input to previous memory and update gate determines how much of the previous state to keep. Update gate in GRU is what input gate and forget gate were in LSTM. We don't have the second nonlinearity in GRU before calculating the output, neither they have the output gate.

Choose LSTM if you are dealing with large sequences and accuracy is concerned, GRU is used when you have less memory consumption and want faster results.
Glove encoding method will give you the result of your sentiment when some random input was given.

Below is all model's graph between loss and their accuracy.
Note: - For every graph blue line refers to the loss and orange line refers to the accuracy.

1) LSTM

	Accuracy
Training Accuracy	93.92%
Testing Accuracy	86.70%

Figure 4.1.1

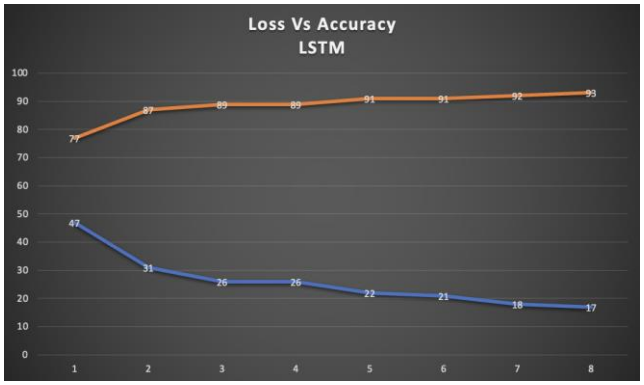

Figure 4.1.2

2) LSTM with CNN

	Accuracy
Training Accuracy	97.16%
Testing Accuracy	87.52%

Figure 4.2.1

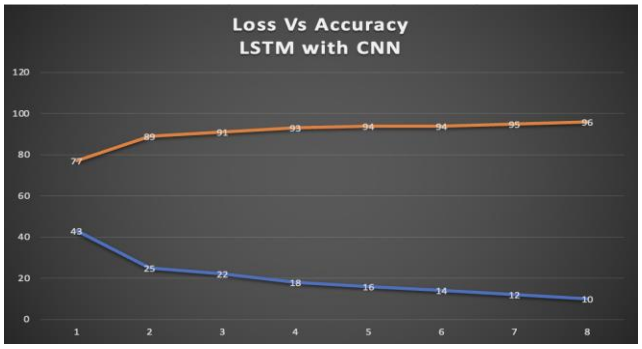

Figure 4.2.2

3) GRU

	Accuracy
Training Accuracy	94.10%
Testing Accuracy	85.42%

Figure 4.3.1

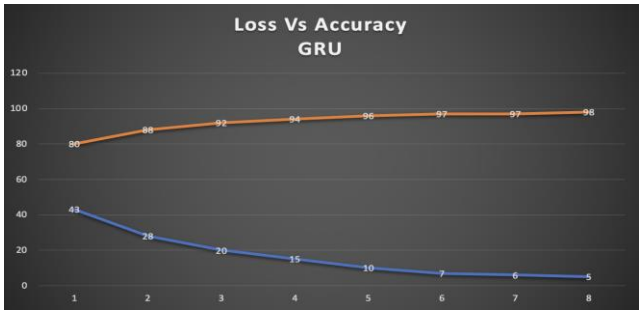

Figure 4.3.2

4) GloVe Encoding

	Accuracy
Training Accuracy	84.31%
Testing Accuracy	83.86%

Figure 4.4.1

Figure 4.4.2

Hence, we can see that for 8 epochs the LSTM with CNN perform better than plain LSTM and GRU. When it concerned with the small memory GRU will be best model suited and when there is a long memory LSTM will be best.

From the reference paper, they have obtained the highest accuracy of around 87.9% for LSTM approach and when they implemented by only CNN approach, they can be able to achieve around 90.23% followed by the GRU can be able to obtain 86.78% accuracy.

Analysis

1. What did I do well?

In our paper, we have used LSTM, LSTM with CNN, GRU and GloVe encoding models and have considered the complete IMDB dataset. In paper they have not used the complete dataset because of its complexity. Rather than plain LSTM to find the accuracy we have added the CNN layer as an input to LSTM layer, it bit complex to do for the given IMDB dataset and we have achieved the highest accuracy of 87.52% in LSTM with CNN model for 08 epochs which is close to the accuracy mentioned in the reference paper.
Also, we have implemented the Glove Encoding in this project to complete dataset and from the random input we could be able to find its sentiment and able to achieve the accuracy of 83.86% accuracy.

2. What could I have done better?

If we had a faster GPU and greater computational capacity, we could have improved the model's accuracy by changing hyperparameters like learning rate, dropout, pool size and epochs. As we used the activation function ‘sigmoid’ we can also try with some other function like ReLu and some different activation function rather than binary cross entropy and could have check the performance of the model. Along with this we have kept top 5000 new words and max review length as 500 to train our model this also can be changed to see performance of model. Furthermore, model can be more optimized by fine-tuning a model and can reduce the power consumption.

3. What is left for future work?

In future, this same dataset can be implemented by using BERT model and analyze the performances of the models and finalize the best suited model for this project. If we have a good CPU, we can increase the more epoch and can check the accuracy and performance of the model. Also, we can try various methods like Naïve Bayes along with CNN to check is there any performance and result difference between LSTM with CNN. If not for the complete dataset, then we can try new methodologies on 20% of the dataset and compare the results.

Conclusion

Natural language processing, text analysis, emotion detection is all used in sentiment analysis to identify, extract, and analyze emotional states and subjective information. We constructed a variety of models to train on the IMDB dataset and compared the different deep learning models based on this dataset in this research. For the complete IMDB dataset, the findings demonstrate that LSTM with CNN has the greatest test accuracy of 87.52 percent. Moreover, we can change the accuracy by tuning the model and changing the hyperparameter like learning rate, activation function and epoch. Addition of CNN layer on the top of LSTM will give more accurate result as it filters the data. Also, we observed that plain LSTM and GRU has not shown much difference in the performance and accuracy when it compared with time and CPU optimization. So, we can conclude that if it compared with less memory GRU model will be best and when it compared to large memory LSTM model will be best as both shows the accuracy 83.86 and 86.70 on the same dataset.

References

[1] Kamil Topal, Gultekin Ozsoyoglu – "Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews".

[2] Md. Rakibul Haque, Salma Akter Lima – "Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews".

[3] Saeed Mian Qaisar – "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory".

[4] Guoxiu He, Wei Lu – "Entire Information Attentive GRU for Text Representation"

[5] Brandon Joyce, Jing Deng "Sentiment Analysis Using Naive Bayes Approach with Weighted Reviews - a case study".