CSE 464 Soft-Computing Fall 2021 Semester Project: Star rating of text using soft computing algorithm

Md Ismail Hossain

Thursday 9th December, 2021

1 Problem Statement

Submitting online review is a very popular trend now a days. Sometimes people put only the text review and it's hard to read all of them and make an overall idea about the service or place or the object. In this project the objective is to understand the health of the organization or service by using machine learning technique by analyzing the text review.

2 Justification of Model Use

Classifying the text is a classification problem. There are several modelling approach to do that but Fasttext, Recurrent Neural Network (RNN), Long short-term memory (LSTM), BERT (Bidirectional Encoder Representations from Transformers) are the mostly used models in the domain. All the RNN, LSTM and BERT are neural network models and Fasttext use the multinomial logistic regression approach.

In RNN the output from the previous steps fed into next step. RNN mainly used for Sequence classification (Sentiment Classification & Video Classification), Sequence labelling (Part of speech tagging & Named entity recognition) and Sequence Generation (Machine translation & Transliteration). In the RNN modelling there are some memory issue, that's why LSTM models introduced and performs way better than RNN.

We have also implemented BERT model to compare the results. Typically LSTM model could only go either left to right of the sentence or right to left. It could not go both direction. On the other hand BERT model is called bidirectional, which means that it could consider the word from both direction of the sentence. BERT is actually an encoder. Which create a fixed encoding vector for each word in the sentence or for each sentences in some case. The BERT base model create 768 size vector for each word. Then that contextualize word vector feed into a Feed Forward Neural Network model to get the final prediction of the sentence rating. The embedding vector for each word actually done through pre trained Masked language model and that per trained model trained through 2500 million Wikipedia word and 900 million other word.

So, in our project we will implement Fasttext ,LSTM and BERT models to fit the text classification models and recommend the best according to the parameter sets and training data.

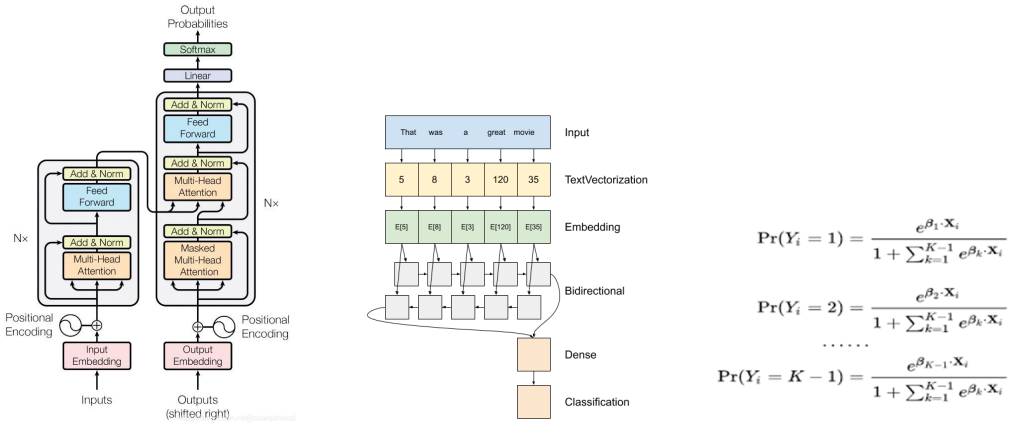


Figure 1: Overview of BERT, LSTM and Fasttext model architecture.

3 Training Data

In this project I would use Yelp review data to train the model. In the training data the response are rating for given review. So, there are 1-5 star, 1 star means worst and 5 star means best.

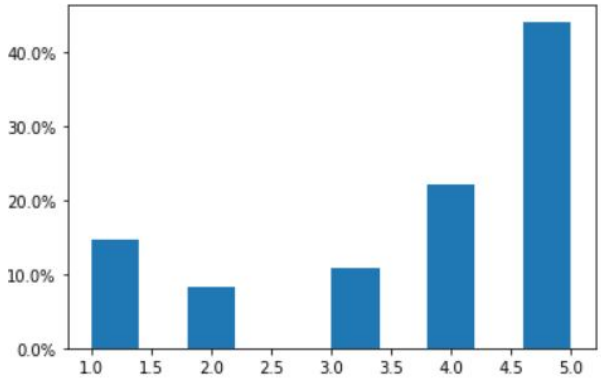
The main data set contain almost 8.6 million of reviews. The average rating for the reviews is around 3.7 star. I used a total of 50k reviews to train the models because of limited computational resource. [Kaggle python instance for data preparation and Google colab for training]

4 Exploratory Analysis

In the original text review rating data we observed that the rating 5 star is highly skewed and the data is highly imbalanced. So, we have to consider equal amount of review for each star in training otherwise the analysis might not be robust or unbiased.

To overcome the computational resource problem and to make the data balance we have selected 10 thousand reviews for each of the rating to train the models. So, we are using total 50 thousand reviews to train the models.

Figure 2: Text rating distribution in training data.



5 Parameter tuning:

We have trained the BERT, LSTM and Fasttext models for couple set of hyperparameter. For the seek of comparison I tried to keep same of the important parameter like learning rate, epoch, validation split, optimizer etc. Inherently there are some parameters which are different in both case. But at least we can make a comparison when learning rate and epoch are similar.

Training setup			
Parameters	Fasttext	LSTM	BERT
Validation Split	0.2	0.2	0.2
Learning rate	0.1	0.1	0.1
Max. word length		280	
Dimension	50		
Embedding size		32	
Optimizer	Adam	Adam	Adam
Activation		Softmax	

The output for both model in terms of accuracy presented in the following plots:

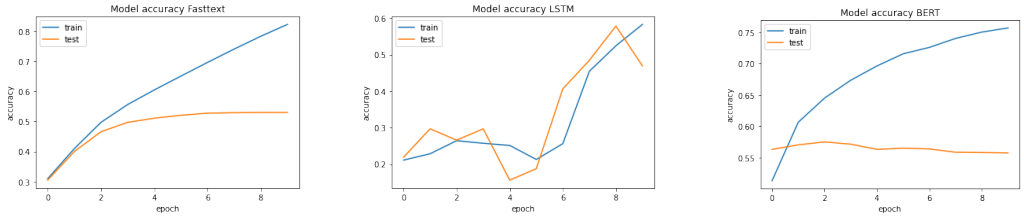


Figure 3: Accuracy after training for Fasttext, LSTM and BERT.

In the training and test accuracy plot we observed a strange pattern between the accuracy over the epoch for training and text set. I suspect a overfitting issue for Fasttext. On the other hand the LSTM model accuracy is not very good, it's lower than 60% even for epoch=10. So, we need to feed more data in the model to feed and tune the parameters.

So, we need to tune the model using more data set and for different sets of parameters to recommend the best one.

6 Conclusion:

It's really very hard to pick the winner among these three models. Because the prediction depends on lots of facts and we observed that non of them are very strong at least for this set of data and sets of parameters. But at least the Fasttest training accuracy is highest and the it's easier to train the model because no previous data preparation is mandatory for this model.We could compare them using other data sets and might get a strong evidence to pick the best.

References

[1] Amjad Abu-Rmileh. How does fasttext classifier work under the hood? https://towardsdatascience.com/fasttext-bag-of-tricks-for-efficient-text-classification-513ba9e302e7, 2019.

[2] Shukhrat Khodjaev. Application of rnn for customer review sentiment analysis. https://towardsdatascience.com/application-of-rnn-for-customer-review-sentiment-analysis-178fa82e9aaf.

[3] Matthew Mayo. Tokenization and text data preparation with tensorflow keras. https://www.kdnuggets.com/2020/03/tensorflow-keras-tokenization-text-data-prep.html, 2020.