Classification of Textual Data: Naive Bayes and BERT

Brandon Park

Sean Li

Yanis Mouazer

Editor: Brandon Park, Sean Li, Yanis Mouazer

Abstract

In this project, our objective was to implement a naive Bayes from scratch and BERT with pre-trained weights and compare the performance of these two algorithms on an IMDB review dataset. We found that the BERT model showed a better performance but was significantly slower to train when finding its benchmark model. The two algorithm used different ways to transform the unstructured text data into numerical features. Most importantly, we were able to distinguish the benefits and costs of deep learning and traditional machine learning methods examining the attention matrix, specifically where the improvements from fine-tuning BERT can produce even greater results on top of its pre-trained corpus. Predicting the sentiment of BERT on its benchmark dataset proved to be more timely and costly than our naive Bayes. Although the findings do not take into consideration the subtle complexities and nuances that exist in any individual's dialogue, our accuracy using both models performed well highlighted the extremities of emotions that can be conveyed in speech.

1. Introduction

Predicting the sentiment of IMDB reviews using the training and testing set helped indicate the benefits and consequences of both traditional machine learning and deep learning NLP techniques. While the complexities of human speech cannot be truly classified with either of our models, both naive Bayes and BERT used these data points to identify the similarities and frequencies of certain words being used in each review. We first compared the performances of these two models, and found that BERT outperformed our naive Bayes implementation although the larger time complexity makes us ponder the trade-off for a marginal improvement in performance. We further delved into the performance of both models, changing the threshold of the frequency to determine if the distribution of reviews were polar in nature. Additionally, fine-tuning our BERT changed how we utilised it for future experiments (Senaratne, 2021), determining the best batch size and epoch when compiling and fitting the model of the neural network. The importance then lay in pre-training on an external corpus of BERT to eventually examine the attention matrix between the words and class tokens (Kokab, 2022).

2. Dataset

The IMDB dataset consisted of 50000 movie reviews split in half for the training and testing set. Each data in this dataset consisted of the movie review itself, and then the sentiment of the review, classified into either "positive" or "negative". To help process the data, we simplified each attribute by vectorizing the sentences of each review into words and labelling "positive" and "negative" into "1" and "0" respectively. Using each word as an element of a vector allowed us to replicate a dictionary-like collection and understand which words consisted in each review. When processing the data for our BERT model, BERT uses a tokenizer process for each review that "tokenizes" some words into a specific token. We only used 512 initial tokens for each review. After the processing, the sentiment of IMDB reviewers were then classified by implementing the naive Bayes and the pre-trained Bert classifier to determine which algorithm would output the most accurate results.

3. Results

Conducting various experiments on the performance metrics of Naive Bayes and BERT led us to conclude that BERT was a clear improvement over the traditional machine learning techniques. Without any additional experiments to see which parameters would be best fit for the two models and relying on its baseline, the report concluded that BERT was the winner in classifying the IMDB reviews (Figure 1).

	precision	recall	f1-score	support
0	0.86	0.84	0.85	12500
1	0.85	0.86	0.85	12500
accuracy			0.85	25000
macro avg	0.85	0.85	0.85	25000
weighted avg	0.85	0.85	0.85	25000

	precision	recall	f1-score	support
0	0.77	0.89	0.83	12500
1	0.87	0.74	0.80	12500
accuracy			0.82	25000
macro avg	0.82	0.82	0.81	25000
weighted avg	0.82	0.82	0.81	25000

Figure 1: Performance of Naive Bayes vs BERT

When training and testing over the full dataset, BERT's accuracy of 0.85 compared to Naive Bayes that achieved 0.82 displayed a significant gap in the language of machine learning. Even without fine-tuning BERT and using only the initial 512 tokens, BERT utilizes previous algorithms and structures that benefit its pre-trained parameters. These parameters that allow BERT to process each review gives it a substantial improvement in accuracy. Having been pre-trained on an external corpus gives it the advantage of understanding the specific task beforehand. The tradeoff with BERT comes with its time complexity. The extraneous pre-training and structure comes with very expensive computational costs that can take very long with such a big dataset like the IMDB classification.

While we felt that the classification reports of the two models was a fair and accurate representation of their performances, we seeked to achieve more optimised models. To understand the distribution of the dataset, more specifically the sentiment of each review and extremities, we scaled different levels of threshold in the Naive Bayes classifier and found these results (Figure 2).

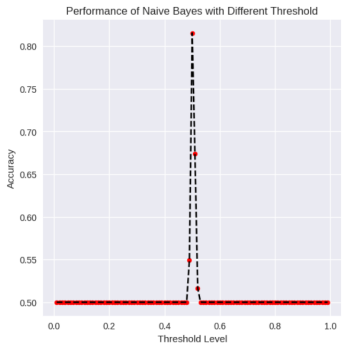


Figure 2: Different Threshold Levels

At a threshold level of 0.5, the accuracy of the model vastly increased, represented in a normal distribution with minimal variance. Changing the threshold level proved that the sentiment of the IMDB reviews were very polar in nature, consisting of either extreme positive or negative reviews, explaining why a little change from the threshold median would effectively misclassify the sentiment of many reviews thus decreasing the performance.

Having visualised the dataset using the Naive Bayes model, we proposed a new way to fine-tune the BERT model. Since BERT was intended for specifically this IMDB classification problem, there were many downstream tasks that required fine-tuning. Due to the sheer size of the dataset, it must be noted that the fine-tuning process was only used with 20 percent of the reviews. This allowed us to maintain our extensive research for various ranges of parameters while not losing the validity and accuracy of the dataset as a whole since BERT was previously trained on an external corpus. We first investigated the number of layers that the multi-layer perceptron was using for BERT on classifying the sentiments.

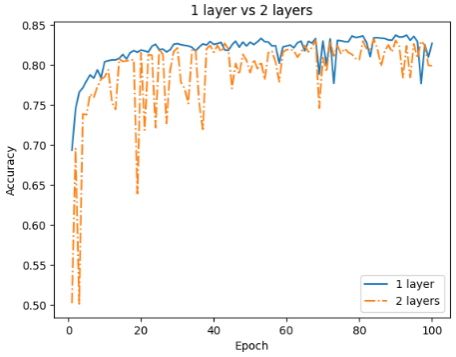


Figure 3: Number of Hidden Layers

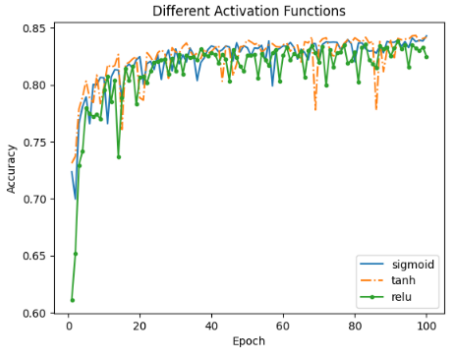
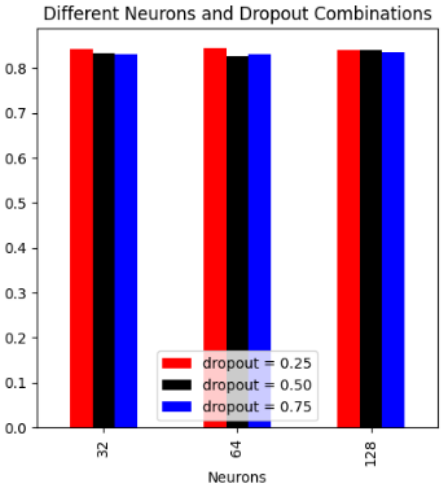


Figure 4: Different Types of Activation Functions

Having 1 hidden layer compared to 2 showed a respective accuracy of 0.83 to 0.81 (Figure 3). Since the dataset that we used to perform this experiment was a fraction of the whole dataset, it is also worthy to note that the accuracy stated below will be lower. However, this would not disprove the notion that 1 hidden layer outperformed because it could reduce overfitting errors that could occur when increasing unnecessary parameters. A binary classification such as the IMDB dataset is a clear example of this situation. Another hidden layer that potentially increases these unnecessary parameters that does not accurately classify could affect our accuracy. Different activation functions used also showed that the sigmoid activation function was the best performer (Figure 4).

Since the output of the sentiment were either classified between 0 and 1, the model can be correctly represented by the function's range. Moreover, the gradient of the tanh and relu function can either be too steep or make some values of the neurons zero, showcased by the steep decrease in both functions at certain epoch stages, harming the model's ability to fit the data as well.



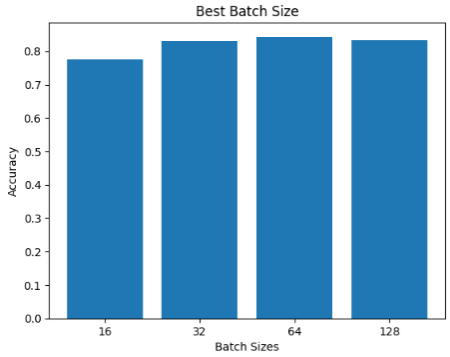


Figure 5: Different Neurons and Dropout Figure 6: Different Batch Sizes

Additionally, we were able to measure that a dropout rate of 0.25 and 64 neurons in the hidden layer was most effective in performance (Figure 5). Through this, we computed that a batch size of 64 would fit the fine-tuned perceptron the best concluding a multi-layer perceptron that would best fine-tune our BERT model (Figure 6).

The Wikipedia pre-trained BERT model has been pre-trained on the text of Wikipedia articles. After our experiments it was found that our baseline BERT model which has been pre trained for predicting IMDB data produced an accuracy of 82 percent whereas the Wikipedia trained model only produced one of 48 percent. By pre-training, the model is able to acquire general language representations from a large text corpus. In turn, this enables the model to pick up on the relationships between words and their meanings as well as sentence and paragraph structure to develop a deeper knowledge of language.

After making a prediction with BERT, we accessed the model's internal state to extract the attention matrices by calling a function to access a specific attribute of the model object. Then we selected the attention matrix for the desired head and block. Finally a heatmap visualisation of the attention matrix was implemented through Seaborn. The rows and columns of the heatmap represent the words in the input sequence, and the colour intensity represents the attention values. For starters, something we noticed is how our BERT model will have a hard time predicting if a review is positive or negative if said review had sarcasm in it. In this review,, "I absolutely loved this movie. The plot was so original and the acting was top-notch. I especially enjoyed the part where the main character fell asleep for half an hour. Truly riveting stuff." producing the heat map for the attention matrix, with head and block number equal to 0.

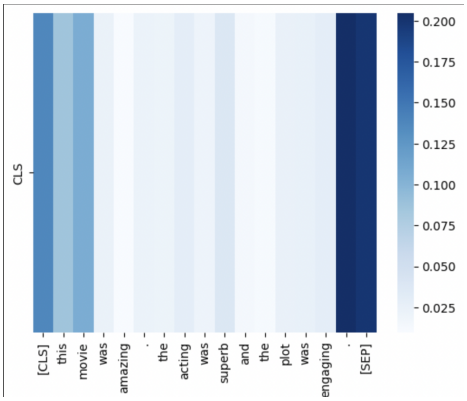


Figure 7: Heatmap of Positive Review

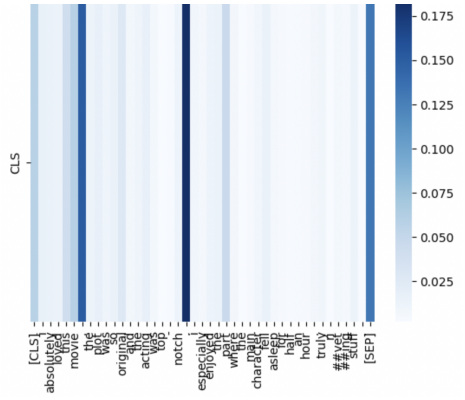


Figure 8: Heatmap of Negative Review

The similarity of both positive (Figure 7) and negative heat maps (Figure 8) shows that the prediction for both were positive. Additionally, conclusions drawn from a single head and block may not provide a complete picture of how the model is making its predictions. It's also important to note that the attention scores are just one part of the model's prediction process. The model uses many other mechanisms, such as its internal weights and biases, to make its final prediction. So even if a period receives a high attention score, its impact on the overall sentiment prediction could be mediated by other mechanisms. The similarities of both figures proves that it was known that a positive review was predicted for both.

4. Discussion and Conclusion

The efficiency of Naive Bayes and BERT models for evaluating the sentiment of IMDB movie reviews was contrasted in this experiment. Our results showed that the BERT model beat the Naive Bayes model in terms of accuracy and F1-score. This demonstrates that the BERT model's ability to understand contextual linkages between words and its pre-training on a large corpus of text make it an effective tool for sentiment analysis jobs. However, the Naive Bayes model required less processing resources and trained more quickly. The model to adopt may depend on the unique requirements of the work, but overall, our findings indicate BERT modelling as a good tool for sentiment analysis of movie reviews. For future exploration, in our attention matrix analysis, it was noted that our models do not perform as well when it comes to detecting sarcasm in a review. To begin with, we can fine tune on a sarcasm dataset. This would allow the model to learn specific language patterns and cues associated with sarcastic text. Another option we can do is to add more features to our models. For instance, the use of punctuation, capitalization or emoticons are signs that can indicate sarcasm.

5. Statement of Contributions

Each group member shared the components of the project equally. Each member was designated with a task to work on, and continued this approach when conducting the tests. The write-up was written with our shared expertise on our analysis of the models of Naive Bayes compared to BERT.

6. References

K̂okab, Sayyida. Transformer-based deep learning models for the sentiment analysis of social media data. Array. 2022, April 10. https://www.sciencedirect.com/science/article/pii/S2590005622000224

Senaratne, Ravindu. Fine-tuning bert for sentiment analysis. Medium. 2021, December 27. https://heartbeat.comet.ml/fine-tuning-bert-for-sentiment-analysis-d59cfad79ff1