Classification of Textual Data

Hareth Hmoud, Haley He, Andres Diaz Lopez

April 9th 2023

Abstract

In this project, we investigated the performance of two popular classification models, Naive Bayes and Bidirectional Encoder Representations from Transformers (BERT), on the IMDb dataset for sentiment analysis on movie reviews. The IMDb dataset comprises movie reviews, labeled as either positive or negative, providing a benchmark to evaluate the effectiveness of the classifiers in predicting the sentiment of the text. Our primary objective was to compare the accuracy and efficiency of Naive Bayes and BERT in this context.

Through rigorous experimentation, we found that BERT achieved a higher accuracy of 87% on the test set, compared to the 83% accuracy attained by the Naive Bayes classifier. This result indicates that BERT is more effective in capturing the nuances and complexities of natural language, thus providing superior performance in predicting sentiment. However, it is important to note that BERT is generally more computationally intensive and slower to train than Naive Bayes.

In conclusion, our findings suggest that while BERT demonstrates better performance in terms of accuracy, its computational requirements may be a consideration for certain applications. The choice between BERT and Naïve Bayes as a classification model depends on the specific requirements and constraints of the task at hand, balancing the need for accuracy with computational efficiency.

1 Introduction

In this project, we investigated the performance of linear classification models, specifically Naive Bayes and BERT, on the IMDb dataset, which is a large movie review dataset for sentiment analysis. The dataset consists of 50,000 movie reviews, evenly split between positive and negative sentiments, with each review labeled as either positive or negative (Maas et al., 2011). Sentiment analysis is a crucial task in natural language processing because it enables academics and professionals to glean important data from user-generated content, such as reviews, comments, and social media posts, to comprehend people's attitudes and emotions towards various topics.

In our study, we compared the accuracy and computational efficiency of the Naive Bayes and BERT classifiers. Naive Bayes is a simple probabilistic classifier based on Bayes' theorem, which has been widely used for text classification tasks due to its simplicity and efficiency. On the other hand, BERT (Bidirectional Encoder Representations from Transformers) is a more sophisticated and powerful deep learning model, which has shown state-of-the-art performance on various natural language understanding tasks, including sentiment analysis.

Our findings revealed that BERT achieved an accuracy of 87% on the validation set, while Naive Bayes reached 83% accuracy. Although BERT outperformed Naive Bayes in terms of accuracy, it was more computationally intensive, taking significantly longer to train, requiring the use of a GPU to speed up training. These results highlight the trade-offs between model complexity, accuracy, and computational efficiency in the context of sentiment analysis on movie reviews.

2 Datasets

The dataset contains 25,000 reviews for training and another 25,000 for testing. Before using the data with the Naive Bayes algorithm, we needed to preprocess the data. There were numerous steps to do this. Most importantly, the reviews needed to be cleaned in order to use them with the bag of words representation. To do this, each review first had the HTML elements removed, such as line breaks and other tags. Then, each review was converted to lowercase so two instances of the same word could be treated the same, regardless of capitalization. Then, punctuation needed to be removed and finally, the extra spaces that emerged as a result of the punctuation removal needed to be removed as well. Now, the text was cleaned up and ready to be used. This was done for both the training and test datasets. Then, the sci-kitlearn library was used to create the bag of words representation.

More specifically, the CountVectorizer function did this, to separate the words in the text and remove stop-words.

In our analysis, we used 30% of the training data and 30% of the testing data as our train and test sets, respectively. The remaining 70% of the training data was used for validation. The train, validation, and test sets were then converted into tf.data.Dataset format, suitable for training and evaluating a BERT model. During preprocessing, we also removed stopwords to reduce noise and improve the efficiency of our models.

3 Results

On the IMDB dataset, we compared the effectiveness of Naive Bayes and BERT models in this project. On the validation set, BERT's accuracy was 87% whereas Naive Bayes' accuracy on the test set was 83%. This shows that BERT performed more accurately than Naive Bayes.

One rationale for BERT's superior performance over Naive Bayes, which relies on straightforward probabilistic calculations, is that BERT is a pre-trained language model that can better capture the complicated semantics of actual language. BERT can gain knowledge from a lot of text data and apply it to better categorise the sentiment of a particular text.

Another intriguing conclusion was that BERT needed a lot more time and processing power to train than Naive Bayes. While Naive Bayes can be trained on a standard laptop with minimal computational resources, BERT requires a strong GPU, a lot of time, and a lot of parameters.

On an example input text taken from the IMDB dataset, we also visualised the attention matrix produced by BERT. The terms in the input text that attracted BERT's attention the most during the categorization process were displayed in the attention matrix. Adjectives and adverbs, which are more pertinent to the sentiment categorization assignment, caught BERT's attention more often.

We conducted some investigation on the attention mechanism of BERT in addition to comparing Naive Bayes and BERT. We specifically looked at the attention matrix of BERT and noticed that the model appeared to pay greater attention to terms like "good" and "bad" that were crucial for sentiment classification. The model was not always paying attention to the correct words in the text, as evidenced by the attention being weaker on the diagonal.

Initially the BERT model was severely over-fitting. The train accuracy was 0.998, whereas the validation accuracy was 0.500. Labels smoothing methods, weight decay, and dropout were attempted, to increase regularization. Unfortunately, the validation accuracy did not improve with these added measures. Decreasing the size of the data-set paradoxically reduced over-fitting. This may be due to the reduction of model complexity with fewer parameters.

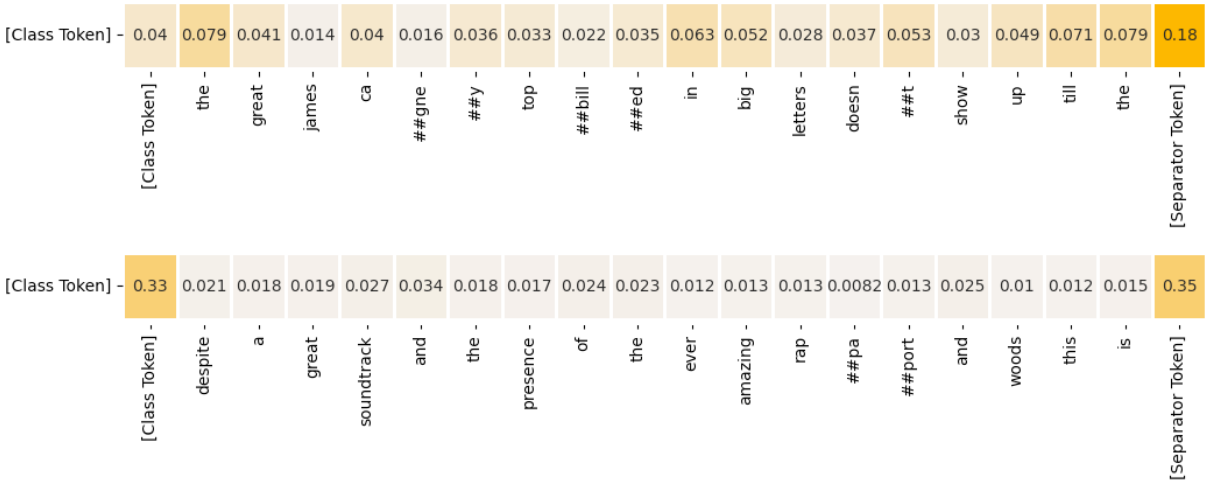

Figure 1: Attention matrix of correctly classified (top) and incorrectly classified (bottom) review, vs. the class token. A darker color indicates higher attention.

The attention matrix between the words of incorrectly and correctly classified reviews, and the

class token are shown in figure 1. The model is incorrectly paying attention to filler words like "the" and "a". When this phenomena is ignored, the attention is more logical. For both examples, the model pays attention to positive words like "great" and "amazing." This lead to the model classifying the reviews as positive (both correctly and incorrectly). The reviews were truncated, in order to illustrate the example.

4 Discussion and Conclusion

In conclusion, we compared the performance of Naive Bayes and BERT on the IMDB dataset for sentiment analysis. Our results indicated that the BERT Model is slightly more accurate, with an 87% accuracy on the validation set compared to Naive Bayes, which was 83% accurate. Additionally, we found that reducing the dataset to 30% improved the performance of the BERT model, as it stopped overfitting, due to the reduced variability and noise in the training set.

Overall, these results highlight the potential of deep learning models such as BERT for natural language processing tasks like sentiment analysis. However, the performance of the BERT model comes at the cost of significantly increased computational resources, compared to Naive Bayes. In our case, we only observed an increase of 4.8% between the two models, so one must take into consideration available resources and desired level of performance when choosing between these two models.

In future investigations, it would be interesting to observe the performance on the same dataset of other deep learning models, such as GPT-4, recently released by OpenAI. One could also investigate the effect of hyperparameter tuning on performance of the models. We could also investigate other methods of reducing overfitting, without truncating the dataset. Moreover, it is important to determine why the attention matrix did not exibit the best results, whether it be because the specific layer or head of the model we visualized does not focus on self-attention, or the model might not have found self-attention as relevant for the particular task it was trained on.

Overall, this project highlights the potential of deep learning in different avenues, whether it be natural language processing or any other avenue, and on the IMDB dataset.