MiniProject 3: Classification of Textual Data

Ethan Pirso, Logan Labossiere, Minh Quan Hoang

COMP 551

Abstract

In this project, we were tasked to implement a Naive Bayes model from scratch as well as a Bidirectional Encoder Representations from Transformers (BERT) with pretrained weights and to compare the results from training the IMDb review dataset. By building the Naive Bayes model completely, it allowed us to gain further understanding of the theoretical concepts by seeing their practicality. By comparing it to BERT (a deep learning model) with pretrained weights which was implemented via a package, the performance difference between deep learning and machine learning is highlighted. This report discusses this difference, the effect of pretraining within this task, and the results for the comparison between BERT and Naive Bayes using multiple comparison metrics.

1 Introduction

The objective of this project was to implement a Naive Bayes model from scratch and Bidirectional Encoder Representations from Transformers (BERT) with pretrained weights and to compare the results of these two algorithms on the IMDb review dataset. The IMDb dataset for binary sentiment classification, which takes reviews and captures similarities between semantic and sentiment similarities between words, must be preprocessed before training uniquely for both models [1]. This dataset allows for complex comprehension of intent from language through vectorization of data and other methods [1]. We were also instructed to implement a Naive Bayes model from scratch and a BERT model with pretrained weights using a package. Afterward, we can feed the preprocessed data to both models to train them. We found that the BERT model had a much higher accuracy than the Naive Bayes model. This is most likely because the Naive Bayes model is not a deep learning model, as there are no neural networks involved, while BERT is a deep learning model. It is also a pretrained language model, which means that it already has a representation of words, which definitely would increase the accuracy of this task involving text. From our results, deep learning as a method seems to offer a higher potential ceiling for accuracy compared to traditional machine learning methods such as Naive Bayes.

2 Datasets

The IMDb Large Movie Review Dataset, which consists of movie review ratings, is a dataset for binary sentiment classification [1]. This means that each review from the dataset is given a rating from 0.0 to 1.0. There are 25000 highly polar movie reviews for training and 25000 for testing. Neutral reviews (scores between 5 to 6 out of 10) were not included.

To preprocess the data for Naive Bayes, we simply removed the HTML tags, the non-alphanumeric characters and whitespace characters, and lowercased all characters. We then obtained the features by fitting and transforming the dataset through a vectorizer. For BERT, we used the BertTokenizer to preprocess the data. Since neural networks work with numbers, it is necessary to represent words using numerical values. We can obtain the word embedding using the encode function from BertTokenizer.

3 Results

Once we trained both the Naive Bayes model and the BERT model, we were able to test them with the test data from the dataset. We trained the BERT model using 3 epochs, a batch size of 16, and a learning rate of $2x10^{-5}$ after trial and error. We also used Adam as the optimizer. After evaluating both our custom-made Naive Bayes model and the BERT model, we achieved an accuracy of 87.16% for the Naive Bayes model and an accuracy of 90.73% for the BERT model, highlighted in the table below.

Naive Bayes	BERT
87.16%	90.73%

The BERT model outperforms the Naive Bayes model on this IMDb review classification task. A part of this reason could be because of the pretraining done with BERT. Pretraining helps with this task in particular because it enables the model to learn more about the general structure of the language, including sentence structure, grammar, and syntax. This pretraining allows BERT to encode the context of each word in a document by looking at the other words in the document. This helps to capture the subtle nuances of language that are important for predicting sentiment.

For the Naive Bayes model, we also investigated the effect of split size on the accuracy. We found that a test/train split of 10/90 performed best (figure 1).

We also computed the attention matrices for select documents from the dataset and plotted them as heatmaps for visualization (figure 2, 3). This was done only for the first five correctly and incorrectly predicted documents to gain an understanding of the output. Figure 1 shows the attention matrices for the first five correctly predicted documents and Figure 2 shows them for the first five incorrectly predicted documents. From the heatmaps, we observed more overall attention was given to correctly predicted documents than the incorrect ones. This is evident in correctly identified documents one, two, and four, where we see lots of attention given to many tokens. On the other hand, only the first incorrectly identified document showed strong attention for any tokens.

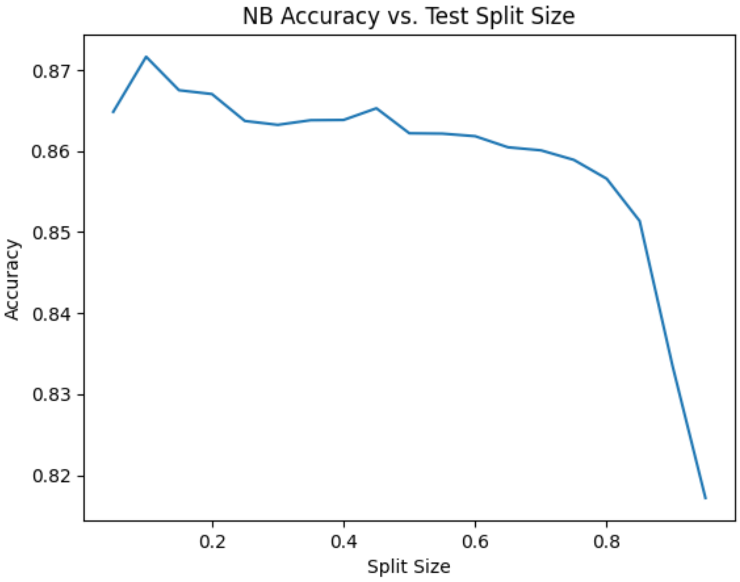
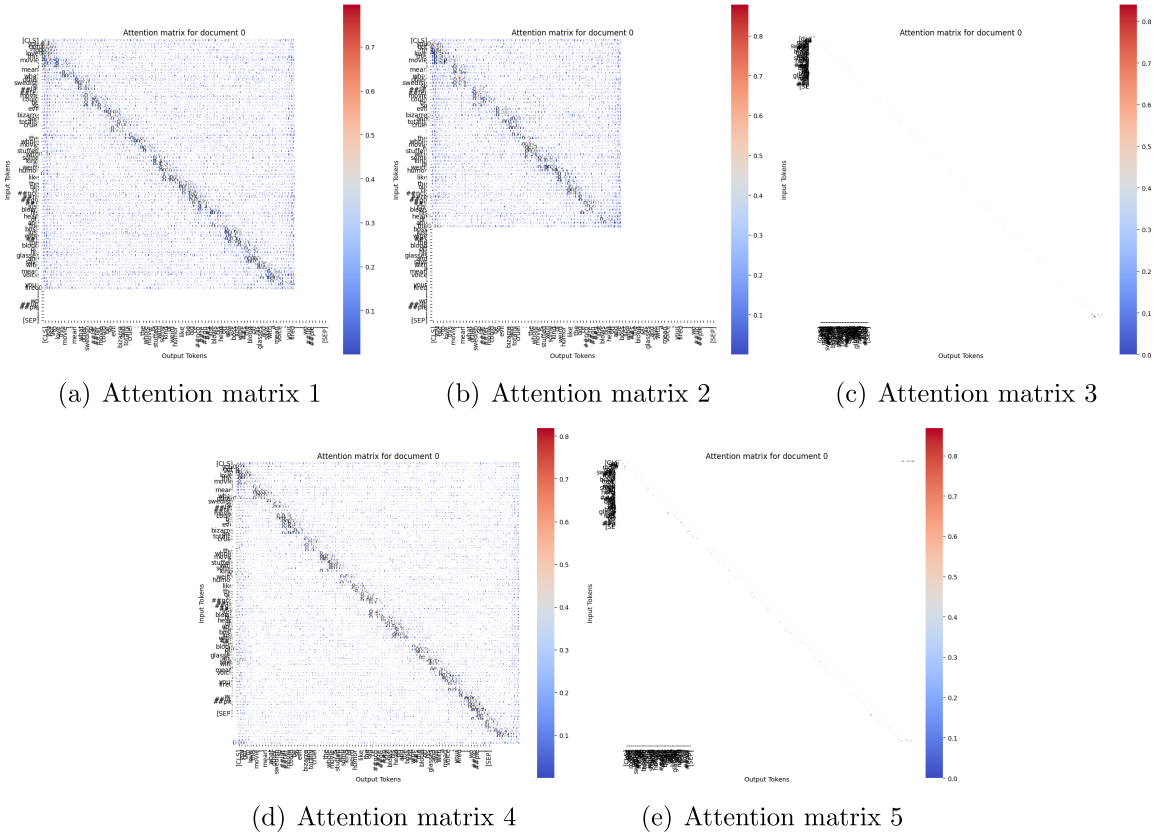


Figure 1: Plot of Naive Bayes accuracy vs. test split size.



(a) Attention matrix 1

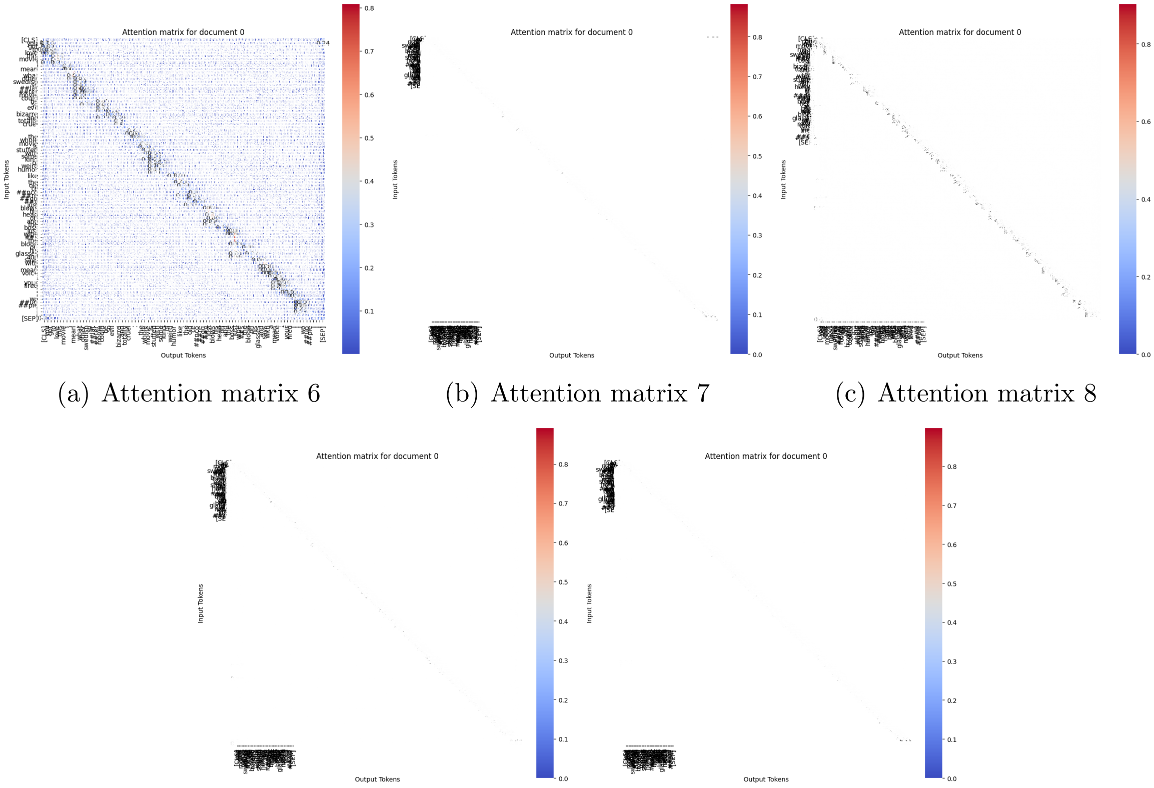
(b) Attention matrix 2

(c) Attention matrix 3

(d) Attention matrix 4

(e) Attention matrix 5

Figure 2: For first 5 correctly predicted documents



(a) Attention matrix 6

(b) Attention matrix 7

(c) Attention matrix 8

(d) Attention matrix 9

(e) Attention matrix 10

Figure 3: For first 5 incorrectly predicted documents

4 Discussion and Conclusion

In conclusion, the BERT model outperforms the Naive Bayes model for this classification task. The accuracy of the BERT model is 90.73% while the accuracy of the Naive Bayes model is 87.16%. We also plotted the heatmaps of the attention matrices for the BERT model for the first five correctly and incorrectly predicted documents. These results suggest that deep learning methods, like BERT, can achieve better performance than traditional machine learning methods, like Naive Bayes. The results also showed that pretraining has a positive effect for this movie review prediction task that involves language and context comprehension. Other avenues to explore on this subject could include exploring how BERT compares to other deep learning methods such as convolutional neural networks (CNN) or recurrent neural networks (RNN) when it comes to sentiment analysis problems such as this IMDb review task. Exploring this could demonstrate what advantages and disadvantages BERT has over those methods and provide a deeper understanding of deep learning.

5 Statement of Contributions

Ethan and Logan were in charge of implementing the models and running experiments on the models. Minh took on the task of writing the report and summarizing the findings from the experiments and design of the models.

References

[1] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).