Binary Classification of Textual Data

Kara Best - 260894988, Niki Mahmoodzadeh - 261045419, Parastoo Gol Mohammadi - 260852345

ABSTRACT

This report details and compares the performance of the Naïve Bayes Model and the Transformer based BERT Model with pre-trained weights, on a binary sentiment classification task. The models were, first, fine-tuned, then trained and tested on the IMDB review dataset. Although the BERT model's prediction accuracy was determined to exceed that of Naïve Bayes by 10%, the Naïve Bayes performance was concluded to be the best suited to the task at hand due to its much greater efficiency.

1. INTRODUCTION

Sentiment analysis is a widely studied topic in the field of natural language processing, where the primary aim is to determine the sentiment of a given document for purposes such as marketing or gauging brand reputation. In this project, we aim to compare the performance of two different machine learning algorithms, Naive Bayes and BERT, for sentiment analysis on the IMDB movie review dataset, using two performance metrics: accuracy and efficiency. The Naive Bayes algorithm, a simple yet effective machine learning method, is implemented from scratch. On the other hand, the BERT (Bidirectional Encoder Representations from Transformers) model, a state-of-the-art deep learning model for natural language understanding, will be utilized with pre-trained weights to explore the benefits of leveraging modern deep learning libraries.

The project consists of three main tasks: acquiring and preprocessing the IMDB data, implementing and fine-tuning the Naive Bayes and BERT models, and running experiments to compare their performances. In the end, we aim to draw conclusions about the performance difference between traditional machine learning and deep learning methods in sentiment analysis tasks and explore the benefits of pretraining on an external corpus, as BERT does.

2. DATASET

2.1 IMDB Reviews

The dataset used in this project is the IMDB reviews dataset, containing 25,000 positive reviews and 25,000 negative reviews with their corresponding classifications. This dataset was evenly divided into the train set and the test set, with 25,000 reviews each, on disjoint movie sets. The labeled train and test sets do not contain neutral reviews, but rather more negative or positive ones. A negative review has a score of 4/10 or lower and a positive review has a score of 7/10 or higher. Furthermore, the number of characters in each sentence seems to vary a lot, as can be seen in the following sample of 2,000 training reviews.

A possible ethical issue with this dataset is that the reviewers never explicitly gave consent for their reviews and their data to be used for research purposes. Researchers need to make sure that they have obtained proper consent from reviewers before using this data for research and training ML Algorithms.

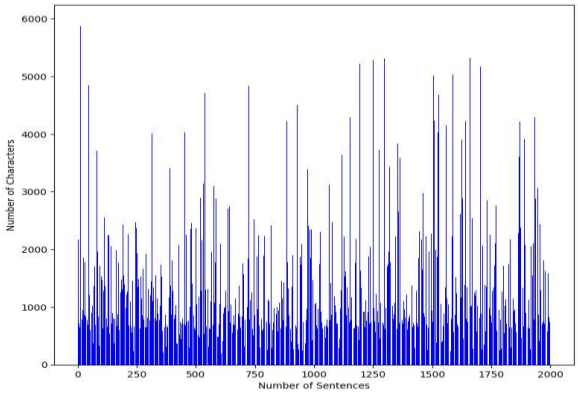

Fig 2.1: Number of characters per sentence

2.2 Data Pre-processing

For the Naïve Bayes model, scikit-learn's CountVectorizer function was used to convert the data from text to a bag of words representation. For the BERT model, the BERT tokenizer function was used to convert the text data to tokens /word pieces.

3. RESULTS

In the following section, experiments performed on the Naïve Bayes model and the BERT model are detailed and their results are reported.

3.1 Naïve Bayes

The experiments using the Naïve Bayes model investigate the effects of excluding certain data from the training dataset during preprocessing on the test accuracy and aim to achieve the highest possible accuracy for a final performance comparison with the BERT model.

The first criterion for word exclusion considered was whether it is present in the scikit-learn's 'english' stop word list. This is a list of English words that generally provide little to no useful information due to frequent use such as "I", "the" or "is". The table below summarizes the test accuracies for no words excluded and for the stopwords excluded, in addition to the vocabulary sizes.

	'english' stop words included	**'english' stop words excluded**
Test Accuracy	82.824 %	82.684 %
Vocabulary Size	74849	74538

Table 3.1: Test accuracy of Naive Bayes for stop word inclusion and exclusion

The reported results indicate that omitting the stopwords from the data leads to lower accuracy. We can conclude that some of the words present in the stopwords list most likely provide valuable context in sentiment classification tasks.

The next elements considered were minimum and maximum data frequency limits. Words are excluded from the bag of words vocabulary if the number of times they are used in the training reviews falls under and/or exceeds the limits imposed. The test accuracy was plotted as a function of the minimum and maximum data frequency limits, as shown in the figures below. We can see that the prediction accuracy peaks at 84.412% for a minimum frequency of 110. The vocabulary size at this point is 3557 which is a significant decrease from the original size of 74849. As the maximum data frequency limit is decreased, we see a decrease in accuracy; therefore, this limit will not be imposed.

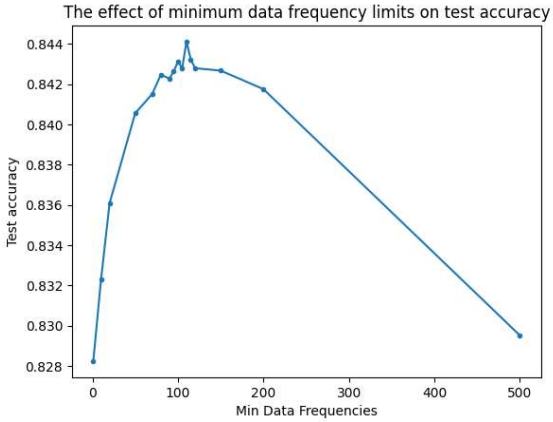

Fig 3.1: Test accuracy as a function of min. df limit

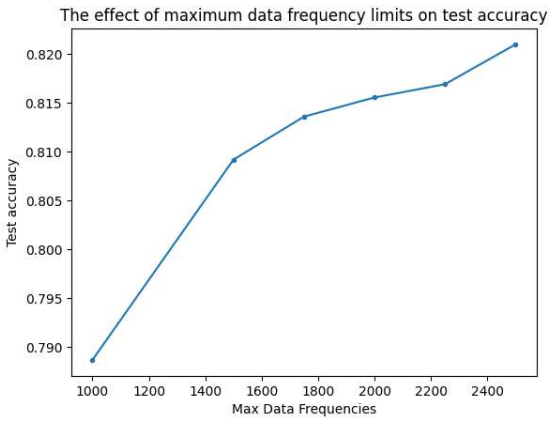

Fig 3.2: Test accuracy as a function of max. df limit

We can conclude that the highest Naïve Bayes model test accuracy was achieved by imposing a minimum data frequency of 110, not imposing a maximum data frequency limit, and including the 'english' stopwords.

3.2 BERT

To evaluate the performance of the BERT model on the IMDB reviews dataset, we used the pre-trained BERT model for sequence classification, specifically the "bert-base-uncased" model. The model was fine-tuned on the dataset for three epochs, and the performance was evaluated on the test set.
The training and evaluation were done using a custom training loop with the AdamW optimizer and a linear learning rate scheduler. The model was trained on a GPU, with the PyTorch deep learning library. The training and evaluation loss and accuracy were tracked and reported for each epoch.
The results are as follows:

Epoch	Training Loss	Test Accuracy
1	0.3092	92.95%
2	0.1620	93.98%
3	0.0722	94.03%

Table 3.2: BERT Model Fine-tuning Performance - Training and Test Loss and Accuracy for Each Epoch

After three epochs of fine-tuning, the BERT model achieved a test accuracy of 94.03%.

3.3 Attention Matrices

The BERT model has 12 layers and 12 attention heads. Below are the attention matrices (from one of the 12 heads) between the words and the class tokens for one correct (Fig 3.3.1) and one incorrect (Fig 3.3.2) classification. As shown in the heatmap of the correct classification for the input text "I really loved this one", the words "really loved" have the highest attention weights, which makes sense for the positive classification. On the other hand, as shown in the heatmap of the incorrect classification for the input text "it is not as awesome as everyone says", the words "it" and "awesome" have the highest attention weights, and the word "not" has a lower attention weight, which can explain this misclassification of the negative review as a positive one.

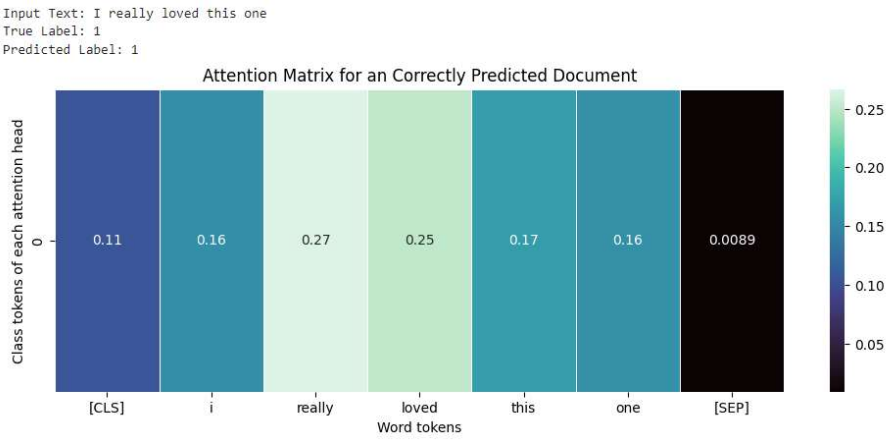

Fig 3.3: The x-axis is the word tokens, and the y-axis is the 0^{th} token, or the class token of an attention head, which is positive in this case.

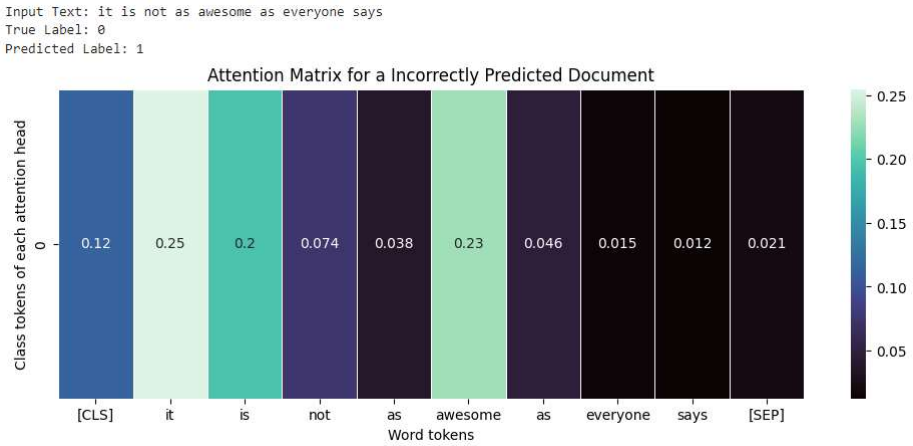

Fig 3.4: The x-axis is the word tokens, and the y-axis is the 0^{th} token, or the class token of an attention head, which is negative in this case.

3.4 Model Performance Comparison

After fine-tuning both the BERT and Naive Bayes models, their performances were evaluated and compared using the metrics test accuracy and execution time. The results are summarized in table 3.3 below and the higher performing model has been highlighted in red.

Observing the results, we notice a tradeoff between accuracy and execution time. The BERT model's accuracy is 10% higher but took a total of 2.6 hours to train and predict, averaging at around 40 mins per epoch. While higher accuracy and receptiveness to fine-tuning is appealing, the context in which the models are being applied must be considered when selecting the most suitable model. The sentiment analysis on a large set of public reviews would most likely be used to continuously gauge and update public consensus on movies. In addition, it is important to note that there are very few ramifications to small inaccuracies in this case. The Naïve Bayes model is able to achieve a high accuracy in less than half a second allowing it to be run continuously with fewer resources while still producing good results. Taking this into consideration, it has been selected as the higher performing model for this sentiment analysis task.

	Naïve Bayes	BERT (3 epochs)
Test Accuracy	84.412 %	94.03 %
Execution Time (s)	0.191	9551.011

Table 3.3: Performance of the Naïve Bayes and BERT models

4. DISCUSSION & CONCLUSION

While more complex deep learning methods such as transformers have generally outperformed traditional machine learning methods like Naive Bayes on sentiment analysis tasks, the specific performance difference can depend on the size of the dataset, the quality of the data, and the complexity of the language patterns involved.

Deep learning methods such as BERT are able to learn more complex and abstract representations of the input data, which can lead to better accuracy on tasks that require a more comprehensive understanding of language. Furthermore, pretraining the BERT model on a large corpus of text enables the model to develop a better understanding of the underlying structure of language (such as syntax and context), which can improve its ability to generalize to new tasks and domains. This is particularly important for accuracy in tasks such as sentiment analysis, where the meaning of a sentence can be subtle and context-dependent, and previous learning can increase the accuracy of the model substantially.

However, in the case where we have smaller datasets with simpler language patterns, traditional machine learning methods like Naive Bayes may perform just as well or even better than deep learning methods. This is corroborated by the high accuracy achieved by the Naïve Bayes model enabled by the low levels of complexity and length of the reviews in the IMDB dataset.

The context in which these models are applied is key when evaluating performance. While the BERT model's accuracy is 10% higher than that of the Naive Bayes model, it took 2.6 hours to produce these results compared to under half a second for the Naive Bayes model. This indicates that the applications of the models differ greatly. Due to the very high accuracy of the BERT model, it is better suited to applications in which there is low tolerance for prediction errors, such as medical testing or for marketing companies needing a precise analysis. In this particular sentiment analysis task, however, there are no devasting implications of false predictions; therefore, the efficiency of the Naive Bayes model while still having high accuracy makes it the higher performing model.

5. STATEMENT OF CONTRIBUTIONS

Kara worked mostly on the Naïve Bayes model, Niki worked mostly on the BERT model, Parastoo worked mostly on the Attention Matrices.

References

1. “Sentiment Analysis.” Ai.stanford.edu, ai.stanford.edu/~amaas/data/sentiment/.

2. “PyTorch.” Www.pytorch.org, pytorch.org/hub/huggingface_pytorch-transformers/.

3. “BERT Testing on IMDB Dataset : Extensive Tutorial.” Kaggle.com, www.kaggle.com/code/atulanandjha/bert-testing-on-imdb-dataset-extensive-tutorial. Accessed 24 Mar. 2023.