Fine-Tuned BERT for the Detection of Political Ideology

Stanford CS224N Custom Project

Alexandre Simoes
Management Science and Engineering Department
Stanford University
asimoes@stanford.edu

Maria del Mar Castaños
Department of Statistics
Stanford University
mmca@stanford.edu

Abstract

In this paper, we apply Bidirectional Encoder Representations from Transformers (BERT) to detect political ideologies in congressional debate transcripts from 2005. For this task, the Ideological Books Corpus (IBC) data set was used, which contains 4,062 sentences annotated for political ideology. Using fine tuned BERT the accuracy achieved was 68%. The F1 Score achieved was 65%, which is more than 30 percentage points higher than former implementations.

1 Introduction

We use BERT in order to detect political ideologies in congressional debate transcripts from 2005 and test its performance for this task. Our interest in this topic comes from its relevance; within many of the issues discussed by politicians word choice entails choosing an ideological position (Iyyeer et al., 2014). Moreover, we want to observe how well BERT executes the task of detecting politicians' transparency in regard to their de jure political standing for further research and testing in this topic.

The data set used for this project is called the Ideological Books Corpus (IBC), which was based on publicly available US congressional floor debate transcripts from 2005. In particular, the data set consists of 2025 liberal, and 1701 conservative biased sentences. Authors have expressed the challenges posed by such data set in text classification tasks. Thus, we implement BERT in order to achieve better results than previous models.

Even though there is research in this topic, the models implemented have achieved an F1 score less than 0.5 (Basak & Misra, 2017). In that sense, we fine tuned BERT in order to achieve the maximum accuracy and F1 score and outperformed previous models used for this task.

2 Related Work

Using Natural Language Processing in order to recognize different political ideologies has been explored by Iyyeer et al. (2014), who used a Recurrent Neural Network to asses this task, the accuracy achieved was close to 0.75. Similarly, Basak & Misra (2017) used a LSTM network for two data sets, one of which was the IBC data set and the F1 Score achieved was around 0.3. In that sense, it was expressed by the authors that training networks on the IBC data set is a challenging task.

Results above are not unprecedented, previous authors have used RNN and bidirectional Rvnn for text classification of opinions. The best F1 score obtained is in the range of $0.5 - 0.6$ (Isroy, et al., 2013).

Devlin, et al. (2018), claim that using Bidirectional Encoder Representations from Transformers to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers outperforms previous Natural Language Processing methods for a wide range of tasks, including text classification.

Sun, et al. (2020) explored Bert's performance on text classification tasks **??**. They propose pre training BERT within the particular task and then fine tuning the parameters within such task as well. They find, that BERT outperforms previous metrics.

3 Approach

BERT (Devlin et al., 2018) is a contextual model that captures relationships in a deep bidirectional way. It allows for the representation of tokens that are dependent on the other tokens in the text. The pre-training tasks, combined with the model architecture, allows BERT to detect semantic relationships between words and sentences that are useful for a variety of tasks. As shown in the original paper (Devlin et al., 2018), when it was released, BERT was capable of achieving state-of-the-art results in common NLP tasks such as GLUE and SQuAD. For these reasons, we chose to fine-tune BERT to detect political ideologies in congressional debate transcripts from 2005 using the IBC data set.

BERT is pre-trained on large corpora for the tasks of next sentence prediction and language modelling. For the second task, BERT uses an innovative approach of considering both the tokens to the left and the right of the masked token when making predictions. The model is open-source and we follow Sun et al. (2020) to fine-tune the pre-trained version of it and the additional untrained classification layer on our specific task. Since the data distribution of our data set may be different from that of the training data set, we further pre-train BERT, with the same tasks mentioned previously, on the IBC data set.

We implement the BERT pre-trained model that contains an encoder with 12 Transformer blocks, 12 self-attention heads, 12 output layers, and the hidden size of 768 ('Bert Base Uncased'). It takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. For each encoder, there is a layer at the beginning, followed by a Transformer architecture for each encoder layer: Attention, Intermediate, and Output. At the end, we have the Classifier layer.

For text classification tasks, BERT takes the final hidden state h go the special token '[CLS]' as the representation of the sequence. Thus, a classifier is added on top of the model to predict:

$$P(Label = c|h) = Softmax(W \cdot h)$$

Where W is the task specific parameter matrix.

For our task, BERT's parameters are fine tuned jointly with W by maximizing

$$Log(P(CorrectLabel = c|h))$$

For training and testing we used the code of Sun et al. (2020) for guidance in order to use the Transformers library with PyTorch; however, Sun et al. (2020) used a different library.

4 Experiments

4.1 Data

As previously mentioned, we used the Ideological Books Corpus (IBC) data set, which was based on publicly available US congressional floor debate transcripts from 2005. In particular, the data set consists of 2025 liberal, and 1701 conservative biased sentences which were picked to be most expressive of political sentiment and manually labeled by a majority voting scheme.

This data set has been proven to be challenging when performing text classification tasks with it as it is manually labeled by a voting scheme which can be subject to subjectivity rather than objectivity.

4.2 Evaluation Methods

We use F1 Score and Accuracy to evaluate the performance of our model for text classification. The F1 Score is the harmonic mean between precision and recall:

$$\text{F1 Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

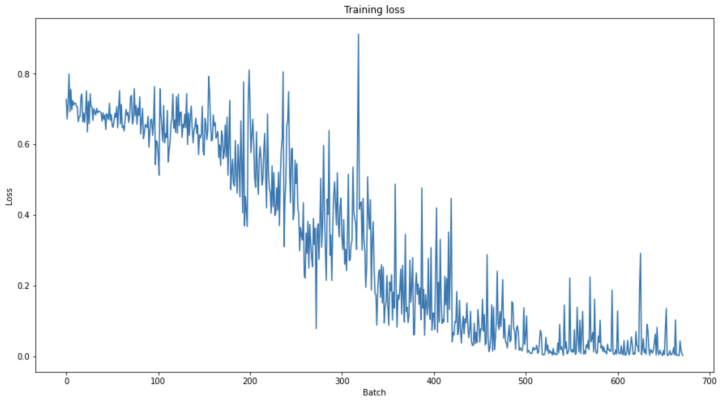
and

$$\text{Accuracy} = \frac{TruePositives + TrueNegatives}{AllSamples}$$

These metrics are usually applied for classification tasks with multiple labels. Additionally, the models against which the performance of BERT is being compared uses these.

4.3 Experimental Details

We first trained the normal BERT model using our dataset for the pre-training tasks of language modeling and next sentence prediction. With an added single linear layer on top of BERT, we then trained the model for classification on 80% of the data set without doing layer selection. An Adam optimizer was used to tune the parameters and the learning rate was $1e^{-5}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, with a warmup rate of 0.1, and L2 weight decay of 0.01. We trained over 8 epochs with a batch size of 32, and achieved the following training lost convergence:



Moreover, in trying different learning rates we achieved the following results:

Learning Rate	Warmup Rate	Epsilon	Batch Size	Accuracy	F1 Score
e^{-5}	0.3	e^{-08}	32	0.60	0.56
e^{-5}	0.1	e^{-08}	32	0.65	0.64
e^{-5}	0.3	e^{-12}	32	0.65	0.60
e^{-5}	0.1	e^{-12}	32	0.68	0.65
$2e^{-5}$	0.3	e^{-08}	32	0.32	0.46
$2e^{-5}$	0.1	e^{-08}	32	0.60	0.61
$2e^{-5}$	0.3	e^{-12}	32	0.63	0.62
$2e^{-5}$	0.1	e^{-12}	32	0.62	0.63

For every combination of parameters, the F1 Score is higher than that of previous implementations for this task and the accuracy is very close to previous values. Moreover, by tuning the parameters, rather than using the default model, we are able to improve performance.

The non-linear activation function in the encoder and pooler of the model was set to "$gelu_new$" as this method outperforms all the others.

Our selection of the parameters was based on the F1 Score and Accuracy Rates.

4.4 Results

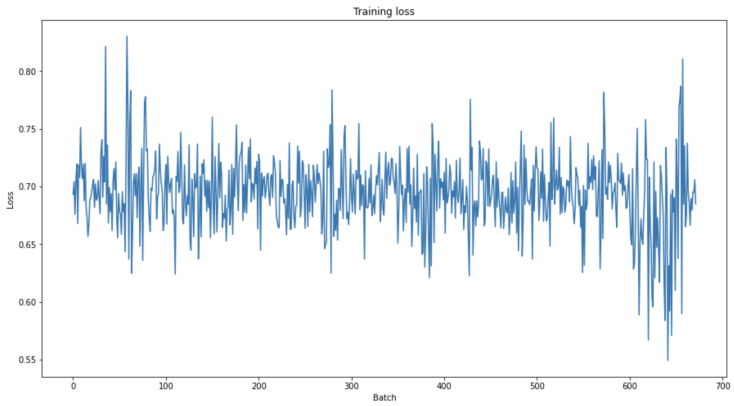
We tested the model in 20% of the data set and achieved an accuracy of 68%. The average F1-score for the two classes was 65%. The highest accuracy achieved by previous models (RNN and LSTM) was 69.3%; however the best F1 score achieved so far is close to 0.4.

Given that our data set is unbalanced and contains large sentences, accuracy is close to what was expected. The F1 score, however, is higher than what we anticipated as we expected it to be closer to what was previously observed in the literature.

Even though BERT doesn't outperform previous models' accuracy, it significantly outperforms them when looking at the F1 Score. This is worth paying attention to as the F1 Score is a better metric when there are imbalanced classes as it is our case.

5 Analysis

For a learning rate of $2e^{-5}$, $\epsilon = e^{-08}$ for the normalization layers, and warmup rate of 0.3, the model fails to converge, and performs very poorly:



As seen in the table above, for every other combination of the parameters the model is able to converge in 8 epochs and performs relatively well. Thus, with a value of e^{-08} for the normalization layers a learning rate of $2e^{-5}$ don't make BERT overcome the catastrophic forgetting problem. In particular, we found that the lower the value of ϵ the lower the value of the learning rate needed to achieve convergence.

We also found that the as the dropout probability for the hidden and the attention layers, the higher the number of iterations needed to achieve convergence.

6 Conclusion

We fine tuned BERT to perform a text classification task over the IBC data set. We found that, regardless of the challenges posed by the IBC data set, BERT is able to outperform previous metrics, such as RNN and LSTM. We are able to replicate Sun, et al. (2020) findings in that an appropriate learning rate, BERT can overcome catastrophic forgetting.

Future works may include a higher dropout probability in order to reduce the possibilities of over fitting. We noticed, that a higher value of $p_{dropout}$ meant a higher number of epochs to reduce the training error. Due to memory constraint this paper does not contemplate this issue; however, it concentrates on BERT's classification performance for this task.

7 References

[1] - Sun et al., 2019. https://arxiv.org/pdf/1905.05583.pdf

[2] - Vaswani et al., 2017. https://arxiv.org/pdf/1706.03762.pdf

[3] - Devlin et al., 2018. https://arxiv.org/pdf/1810.04805.pdf

[4] - Wu et al., 2016. https://arxiv.org/abs/1609.08144

[5] - Iyyer et al., 2014. https://www.aclweb.org/anthology/P14-1105.pdf

[6] - Chen et al., 2017. https://www.ijcai.org/Proceedings/2017/0510.pdf

[7] - Code - https://colab.research.google.com/drive/1ywsvwO6thOVOrfagjjfuxEf6xVRxbUNO

[8] - Dataset source - https://people.cs.umass.edu/ miyyer/ibc/. Accessed in 03/03/2020