Using BERT for Hyperpartisan Classification

Mehdi Drissi and **Vivaswat Ojha** and **Pedro Sandoval**
Harvey Mudd College, CA
{mdrissi, vmojha, psandovalsegura}@hmc.edu

Abstract

In this experiment, we applied the recently developed BERT (Bidirectional Encoder Representations from Transformers) model for language representation to the task of classifying news articles as hyperpartisan or not. This was the challenge for SemEval task 4. To deal with only having a small training dataset of high quality labels we used a semi-supervised approach. Firstly, we used 600,000 source-labeled articles without their labels as a pre-training data set for unsupervised tuning. We pre-trained the model on this data set for 5 epochs. Then, 645 hand-labeled articles were split into a training set, which was 80% of the articles, and a validation set, which was 20% of them. We trained the model for 100 epochs on the training set. On the validation set, different versions of our model had achieved around 80% accuracy. Furthermore, on the test set provided SemEval, we achieved an accuracy of 0.77 against a baseline of 0.46, the second-best result in the competition.

1 Introduction

The problem of classifying data is one that has been studied for decades. This problem has become increasingly salient as data becomes more prominent and information is widely disseminated. In recent times, there has been particular concern about identifying news articles that are misleading to their readers, a phenomenon often associated with ”fake news” distributed by partisan sources to mislead their readers.

In this report, we describe our approach to the problem of classifying whether an article is hyperpartisan using the text of the article. This task is a difficult one due to the ambiguity of the label - it is hard to find a meaningful data set for such a problem without having humans manually label articles to begin with. For this classification task, we are interested in training a BERT model to understand whether an article exhibits blind, prejudiced, or unreasoning allegiance to one party. If an article has any of these features or demonstrates a unreasonable affinity to a cause or person, it is likely the article is hyperpartisan. It is important to note that we do not investigate these features individually, as the BERT model is language representation model that is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers.

The goal for our final project was to use transfer learning to fine-tune a model for the document classification task. The BERT model has been used to learn useful representations for a variety of natural language tasks. Because the model was able to achieve state of the art performance in multiple NLP tasks after being fine-tuned, we show that it also performs well on the hyperpartisan sentiment classification as well. We also use using semi-supervised learning, since we first use unsupervised learning on the set of 600,000 source-labeled articles, then train using supervised learning for the 645 hand-labeled articles. The main rationale for training our model in this fashion was our belief that learning on source-labeled articles would bias our model to learn the partisanship of a source, rather than that of the article. In particular, the accuracy of the model on validation data labeled by article differed heavily when the articles were labeled by publisher. So, only the hand-labeled training data was used for supervised training. The remainder of the training data was used to pre-train the model in an unsupervised manner.

2 Previous Work

We relied primarily on the insights provided the following publications to both familiarize ourselves with the BERT model and understand the underlying structure of the model we're using.

The papers we read are:

- Devlin, Jacob, et al."Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- Vaswani, Ashish, et al. Attention is all youneed. Advances in Neural Information Processing Systems. 2017.
- Dai, Andrew M., and Quoc V. Le. "Semi-supervised sequence learning." Advances in neural information processing systems. 2015.
- Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2010): 1345-1359.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).

The primary paper we relied on for this project is "Bert: Pre-training of deep bidirectional transformers for language understanding.". Its main objective was to develop the language representation model BERT (Bidirectional Encoder Representations from Transformers) that can be easily fine-tuned for many different tasks. We used this fine-tuning feature to create an effective model for our classification task. We used the model in a similar way to the paper's examples but adapted it as necessary to suit our project. The methodology of the paper seems quite sound and it seems logically consistent. Their results show that BERT outperforms the state of the art for a number of NLP tasks that they examine. We were able to adapt BERT to perform similarly well in classifying hyperpartisan articles as well, mirroring the approach the authors took. (Devlin et al., 2018)

In the paper by Vaswani et al., the authors develop a neural network architecture called the Transformer based on attention mechanisms. It is important to our topic because it is one of the papers that contributed to the development of the BERT model we are using. Our work is not very similar to the authors of this paper - while they were also using their model on language tasks, the main importance of this paper for our work is its contribution to BERT. The methodology of the paper makes sense on the whole, especially given it is successfully used for the BERT paper as well. The paper's main results are performing well on certain machine learning tasks with this model. We will make use of its contributions to BERT for our project. (Vaswani et al., 2017)

The paper by Dai et al. examines two approaches for pre-training a neural model on a large amount of unlabeled data to improve performance on a smaller amount of labeled data. One approach is to pre-train the model to be a language model while the other is to pre-train it to be a sequence autoencoder. BERT, the model we are working with, was also pre-trained as a language model. We intend to further pre-train it using the large 600k training dataset. We do not plan to use the 600k with their labels due to the low quality of the labels. This should be helpful as the well labeled data is only around 600. The paper found both pre-training methods to work and they found the sequence autoencoder to perform slightly better. One major distinction in the BERT language model vs the language models they examined is that BERT was a bidirectional language model which led to the need of masking a few tokens during training (essentially filling in blanks). (Dai and Le, 2015)

"A survey on transfer learning" by Pan et al. examined transfer learning in classification, regression, and clustering. The main motivation of reading this paper was getting a broader sense as to the usage of transfer learning to make sure we were not missing any major well known idea useful in other tasks. The paper covered a good deal of ground, but we did not find much useful information not already covered in the other papers.(Pan and Yang, 2010)

The Sennrich et al paper examined a tokenization strategy. You begin by starting off with your dataset as characters. Then you find the most common bigram and combine them into a new word piece, with the caveat that you do not merge pieces with spaces. You repeat this process until you get to a desired vocabulary size. The benefit of this method is common words will be a part of the vocabulary. Uncommon words will be broken up into more meaningful pieces like suffixes and prefixes. We found this technique well-motivated and are currently using it as our tokenization strategy. It is necessary for us to use it as BERT was pre-trained with this tokenization strategy.(Sennrich et al., 2015)

3 Methodology

The main method we are operating with right now is adjusting the BERT model for the data available and evaluating its performance.

We note that, in our work so far, we are focusing primarily on the smaller data set of 645 hand-labeled articles, both for training and for validation. Our approach has been to take the first 80% of this data set for our training set and the last 20% for the validation set. We believe that, given the test set is also expected to be hand-labeled, the 645 articles are much more representative of the final test set our model will be tested on.

Due to an intrinsic limitation of the BERT model, we are unable to consider sequences of longer than 512 word pieces for classification problems, which is significantly shorter than the longest article from the training set, which is around 6,500 words. This means that we have been operating with truncated versions of the articles where we only consider the first 100, 250 or 500 word pieces.

Note that when we refer to word pieces, we are referring to the byte-pair encoding that BERT relies on for tokenization. These can be actual words but less common words may be split into pieces. The details of this process are discussed in ”Neural Machine Translation of Rare Words with Subword Units” by Sennrich et al.

The next method we employed was training the model further through unsupervised training on the larger data set of 600,000 articles with the expectation of increasing our accuracy on the hand-labeled validation set further.

4 Results

There are two main types of accuracy we will report on. One accuracy is doing cross validation on the training dataset that is labeled by publisher, while the other is validation accuracy on the 20% of hand-labeled articles. The two accuracies end up having similar values, but it is important to note models trained for publisher classification perform poorly on the hand labeled articles and vice-versa.

We began by building a baseline Multinomial Naive Bayes model which had a cross validation accuracy of nearly 80 percent. In an attempt to improve this model we tried seeing if the model would perform better if it only used outlines of the text. The hope was the outlines would only contain the key sentences and those sentences should better capture the hyperpartisan nature of the article compared to less important sentences. We extracted outlines by selecting sentences that had the highest average TFIDF score. This summarization technique was a minor tweak of SumBasic (Nenkova and Vanderwende, 2005). Our extension ended up performing slightly worse. This reveals that even sentences that are not central to the article can be beneficial to classification.

Model	Cross-validation Acc
MNB	78
MNB (Outlines)	71

Table 1: Lab 7 and Lab 8 results

For our current model, the results we have are based on the validation set we constructed using the last 20% of the hand-labeled articles. The main results we can report currently are the accuracies of the different versions of the model on our validation set with sequence lengths of 100, 250 or 500. These accuracies are not directly comparable to the prior accuracies given they measure on a different piece of validation data.

Max Seq Length	Validation Acc
100	76.7
250	75.9
500	79.1

Table 2: Validation Accuracy Using Last 20% of Hand-labeled Articles by Max Sequence Length

Next, we pre-trained a BERT model using unsupervised learning. We were curious to see whether a pre-training step before supervised training would result in additional performance benefits. For our dataset, we used all of the articles by publisher and allowed the model to train on these in an unsupervised fashion. We collected similar data to make sure that this model performs worse than training supervised, as can be seen in Table 3.

As the reader can tell, the model that was only trained unsupervised did not perform well. This is reasonable as the classification part of the model, at this point is randomly initialized. However, our next intended step was to continue to train this BERT model on the entirety of our hand-

Max Seq Length	Validation Acc
100	27.9
250	27.9
500	27.9

Table 3: Validation Accuracy On Solely Unsupervised Model Using Last 20% of Hand-labeled Articles by Max Sequence Length

labeled articles. We refer to this model as the Fully Trained Model. We found improvements for shorter sequence lengths (100 and 250) but a decrease in performance for 500 word piece chunks. It is interesting that the accuracy for the 500 word piece chunk was lower than the 250 word piece chunk as the longer chunk should have been more informative. This likely indicates training difficulty from the small dataset and potentially from dealing with long sequences.

Max Seq Length	Validation Acc
100	79.8
250	82.9
500	75.2

Table 4: Validation Accuracy On Fully Trained Model Using Last 20% of Hand-labeled Articles by Max Sequence Length

We then decided to evaluate our model on the actual validation dataset called **articles-validation-20180831.xml**. As expected, our model did not perform very well and achived an accuracy of 57.7% as can be seen in Table 5 confirming that classifying by publisher and hand-labeled are two distinct tasks.

We also were able to evaluate our model on the *SemEval 2019 Task 4: Hyperpartisan News Detection* competition's **pan19-hyperpartisan-news-detection-by-article-test-dataset-2018-12-07** dataset using TIRA. Our team evaluated our results before the early bird deadline and achieved results that placed our model in second place based on our accuracy of 77%. Our model submission used a maximum sequence length of 250. Full results for the test dataset can be found in Table 6.

Accuracy	57.7%
Precision	57.6%
Recall	58%
F1	0.578

Table 5: Metric Summary of Fully Trained Model on the Official Validation Dataset

Accuracy	77.1%
Precision	83.2%
Recall	67.8%
F1	0.747

Table 6: Metric Summary of Fully Trained Model on the Official Test Dataset

Based on our successful performance on the official test dataset for the *SemEval* competition, we decided to train a larger BERT model. Our previous BERT Transformer model was a base model consisting of 12 layers and 110 million parameters. The next model we trained was a larger version consisting of 24 layers and 340 million parameters. We refer to this larger model as the Fully Trained Large Model, given that we trained the model in the same way as the base model (first training unsupervised on articles by publisher, then training supervised on the hand-labeled articles). The model took approximately 3 days to train. Our validation results for this larger model can be found in Table 7. Note that comparing the base BERT Transformer model to the larger model is best done by analyzing Table 4 in comparison to Table 7.

Max Seq Length	Validation Acc
100	72.1
250	72.1
500	72.1

Table 7: Validation Accuracy On Fully Trained Large Model Using Last 20% of Hand-labeled Articles by Max Sequence Length

We note that the performance of the larger model is noticeably worse than the original. This is unexpected as the experiments in (Devlin et al., 2018) consistently found improvements when using the large model. The large model's high capacity may have lead to difficulty in fine-tuning with such a small dataset. Its validation accuracy being consistent regardless of the sequence length used is also suspicious and something we plan to investigate more closely. While it could just be a bug, the code was the same besides changing the model size.

5 Discussion

In this project, we adapted the existing language model BERT for the task of hyperpartisan news classification and trained it on a subset of the hand-labeled articles provided by SemEval. We also experimented with models that were trained over the larger 600,000 article data set in an unsupervised manner. When testing the accuracy of our models on the validation set, we obtained a remarkably high accuracy of around 80%, with some noisy differences between models that made use of different numbers of words from the article. Furthermore, on the test set provided by SemEval, we achieved an accuracy of about 77%, allowing us to achieve a second-place ranking in the competition. Our model substantially outperformed the baseline model proposed by the organizers of the competition, which had an accuracy of around 46%.

These successful results demonstrate the adaptability of the BERT model to different tasks. With a relatively small training set of articles, we were able to train models with remarkably high accuracy on both the validation set and the test set.

In future work, we hope to expand upon this model further and tweak it in ways that could potentially further boost the accuracy. For instance, we would like to experiment with methods that allow us to train a model that makes use of the entire article. This could involve just aggregating over all the 500 word sequences present in a given article. Other ways of improving our test accuracy would involve tricks like hand labeling some more articles ourselves and adding them to the training set or adding linguistic features.

We would also like to explore alternative methods of using the large 600,000 dataset. In particular, while we do not have labels for those articles that we are confident in, we do know that each article should have a consistent label across the entire article. So we would expect the model to give the same label for each chunk of the article. If an article is classified as the same label for most chunks that can act as strong evidence that we can incorporate it into the training set with the appropriate label. We may also want to see if we can use the publisher labels in a way that does not lead to the model focusing too much on them. This could be done by either heavily down-sampling the publisher articles or weighting them to a lower degree in the loss function.

References

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised Sequence Learning. pages 1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Ani Nenkova and L Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research Redmond Washington Tech Rep MSRTR2005101*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (Nips).