Hyperpartisan News Detection

Anmol Singhal
IIIT Delhi
2017332

Chirag Jain
IIIT Delhi
2017041

Tejas Oberoi
IIIT Delhi
2017367

Abstract

Hyperpartisanship refers to a particular text or a news article having an extreme left or right-wing standpoint. Such content tends to show a significant bias towards a specific faction or political party and can be misleading for people at large. Manipulative reports in the print and electronic media play a vital role in determining the reader's opinion. Therefore, it is essential to ensure that media reports do not have any hyperpartisanship associated with them. In this project, we use contextualized BERT embeddings (Devlin et al., 2018) with deep learning models to detect any form of hyperpartisanship in our dataset consisting of news articles. The results obtained show a significant accuracy score on the test set.

1 Introduction

In today's world, sensationalism draws people's attention which might not always help one gain a good insight about the entity in question. Ensuring that the content published in newspapers and online media is devoid of any inherent bias is the need of the hour.

Therefore, in recent times, the primary domain of interest for researchers and linguists has become to detect false information in news media articles, which includes fake news detection, clickbait analysis, and detecting hyperpartisanship. There has been ample research done on the first two problems. Detecting hyperpartisanship in articles, on the other hand, is a relatively new domain and has sufficient scope for research.

To perform the intended task, we use contextualized BERT ((Bidirectional Encoder Representations from Transformers) Embeddings (Devlin et al., 2018)and represent each sentence in our article as a vector. We obtain the sentence vector by taking the average of pre-trained word embeddings of the tokens in each sentence. Each article is then represented as a sequence of these sentence embeddings.

After generating the embeddings for each sentence, we train a Convolutional Neural Network (CNN) with Batch Normalization and to learn a representation of a document and further predict whether it exhibits hyperpartisanship or not.

We consider using the SemEval 2019 Task-4(Kiesel et al., 2019) dataset for carrying out our task. It consists of 2 datasets, including a manually annotated by-article dataset which contains about 645 articles and a by-publisher dataset which is more extensive and consists of about 600K articles. The By-Publisher dataset is constructed using distant supervision.

Performance on our model reveals that using the by-article dataset for training helps us gain sufficient accuracy score, but this score reduces when we train the model on the by-publisher dataset. We conclude that the usage of the by-publisher dataset does not help in improving the performance of our model.

2 Related Work

The first work regarding hyperpartisanship was carried out by (Potthast et al., 2018), who used a stylometric analysis to detect hyperpartisanship in articles. This paper modelled the task of classifying documents based on whether they are hyperpartisan or not; however, it argues that detecting if a hyperpartisan article is left or right-wing is a non-trivial task. This work motivated further research in the field.

The SemEval 2019 Task 4(Kiesel et al., 2019) involved Hyperpartisan News Detection. Forty-nine teams submitted a valid run on the portal and some achieved high accuracy values on the data, which was the primary evaluation metric.

We draw inspiration from the work done by Team (Jiang et al., 2019) who ranked one on the by-article dataset. They incorporate contextualized ElMo(Peters et al., 2018) embeddings and train a CNN on the by article dataset obtaining an accuracy of 0.822 on the test set. The sentence We obtain the sentence ElMoembeddings by taking the average of word ElMo embeddings.

The team which came 2nd, Vernon Fenwick(Srivastava et al., 2019), took a different approach and combined sentence embeddings with more domain-specific features and a linear model. Other top performing SemEval(Kiesel et al., 2019) teams used various models, including generating handcrafted features from our data, SVM, sentiment analysis, stylometry analysis and n-grams.

The idea is to compare our performance on the same dataset with other teams and propose a better methodology to carry out this classification problem.

3 Methodology

3.1 Dataset and Preprocessing

The by-article dataset is a manually annotated corpus containing 645 articles whereas the by-publisher dataset contains 600K articles, compiled using distant supervision.

The datasets are present in XML format. Therefore, we first extract the title and article text from the XML representation and further split paragraphs constituting each article into sentences using Spacy. We do not remove punctuations and special characters in our data and maintain the case. We replace all white space characters with a single space and all numerals with a unique number token.

After performing preprocessing and tokenizing each sentence, the by-article set reveals that the maximum, mean, and minimum numbers of tokens are 6470, 666, 19 respectively. On the other hand, by-publisher dataset statistics are 93714, 796, 10, respectively. A maximum of 512 initial tokens per sentence is used for each sentence embedding, as specified in BERT.

3.2 Model Architecture

The main architecture of our Convolutional Neural Network consists of 2 parallel convolutional networks, and each network has a convolution layer which is followed by batch normalisation and max-pooling. A relu activation function is applied after the output of every convolutional layer. The output activation function used after the dense layer is softmax activation.

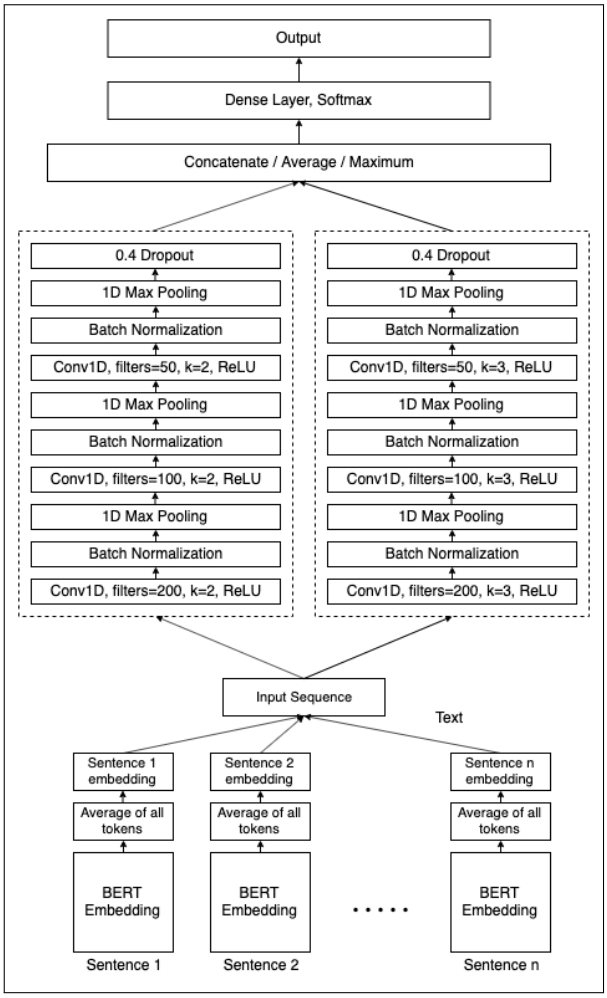

Figure 1: Architecture

3.3 Experimentation

We first implemented a network of two sequential neural networks. Without using batch normalization, we could see that our model was overfitting. Upon applying batch normalization, our validation accuracy improved. Then we found that using two sequential neural networks would mean that we are introducing a dependency in weights of both

models. So as a next step, we moved to a parallel network of neural networks. The architecture is displayed in figure 1.

4 Evaluation and Analysis

We have used a holdout set(20 per cent of by-article dataset) for evaluation. K-cross validation is used with k=10 during training. An 80-20 Train-test split is imposed further during each fold.

4.1 Results

Model	Validation Acc.
Baseline	0.81
CNN with B.N.	0.78
CNN with B.N. & E.L.	0.798

Table 1: Accuracy Scores Obtained

5 Conclusion

The accuracy scores obtained on the validation set by our model surpass the accuracy values obtained by all baselines considered. Thus, using pre-trained BERT embeddings and feeding them to a light-weight Convolutional Neural Network (CNN) helps us achieve performance that is at par with the state of the art. Since CNN has a lot of parameters as compared to the size of the training set, we expect significant variance in the results obtained. Therefore, we take an average of an ensemble of models to get our final predictions. The proposed model tackles the problem of hyperpartisan news detection well and achieves high accuracy on the validation set.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team bertha von suttner at SemEval-2019 task 4: Hyperpartisan news detection using ELMo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Vertika Srivastava, Ankita Gupta, Divya Prakash, Sudeep Kumar Sahoo, Rohit R.R, and Yeon Hyang Kim. 2019. Vernon-fenwick at SemEval-2019 task 4: Hyperpartisan news detection using lexical and semantic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1078–1082, Minneapolis, Minnesota, USA. Association for Computational Linguistics.