Team JACK RYDER at SemEval-2019 Task 4: Using BERT Representations for Detecting Hyperpartisan News

Daniel Shaprin[1], **Giovanni Da San Martino**[2], **Alberto Barrón-Cedeño**[2], **Preslav Nakov**[2]
[1]Sofia University "St Klimen Ohridski", Sofia, Bulgaria
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
shaprin@uni-sofia.bg
{gmartino, albarron, pnakov}@hbku.edu.qa

Abstract

We describe the system submitted by the Jack Ryder team to SemEval-2019 Task 4 on Hyperpartisan News Detection. The task asked participants to predict whether a given article is hyperpartisan, i.e., extreme-left or extreme-right. We propose an approach based on BERT with fine-tuning, which was ranked 7th out 28 teams on the distantly supervised dataset, where all articles from a hyperpartisan/non-hyperpartisan news outlet are considered to be hyperpartisan/non-hyperpartisan. On a manually annotated test dataset, where human annotators double-checked the labels, we were ranked 29th out of 42 teams.

1 Introduction

SemEval-2019 Task 4 (Kiesel et al., 2019) asks to distinguish between articles that are extremely one-sided, i.e., extreme-left or extreme-right, and such that are not. The organizers provided two datasets:

1. **By article:** A small dataset of 645 manually annotated articles (*BA* in the following).

2. **By publisher:** A large dataset of 750,000 articles annotated using distant supervision, where an article is considered hyperpartisan if its source is labeled as such (*BP* in the following). The set is separated into 600,000 articles for training (*BP-train*) and 150,000 articles for validation (*BP-val*).

Furthermore, two test sets, one annotated by article (*BA-test*) and one annotated by publisher (*BP-test*), were hidden from the participants and they were used for getting the final scores for the competition. The task is a binary classification one, where each article is to be assigned one of two possible classes: *hyperpartisan* and *non-hyperpartisan*.

2 Related Work

Media bias was used as a feature for "fake news" detection (Horne et al., 2018a). It has also been the target of classification, e.g., Horne et al. (2018b) predicted whether an article is biased (*political* or *bias*) vs. unbiased. Similarly, Potthast et al. (2018) classified the bias in a target article as (*i*) left vs. right vs. mainstream, or as (*ii*) hyper-partisan vs. mainstream. Left-vs-right bias classification at the article level was also explored by Kulkarni et al. (2018), who modeled both text and URL structure. Some work targeted bias at the phrase or the sentence level (Iyyer et al., 2014), for political speeches (Sim et al., 2013) or legislative documents (Gerrish and Blei, 2011), or targeting users in Twitter (Preoţiuc-Pietro et al., 2017). More recent work has targeted the political bias of entire news outlets (Baly et al., 2018, 2019). Another line of related work focused on propaganda, which is a form of extreme bias (Rashkin et al., 2017; Barrón-Cedeño et al., 2019a,b). See also a recent position paper (Pitoura et al., 2018) and an overview paper on bias on the Web (Baeza-Yates, 2018). Overall, most of the above work focused on finding effective representations, e.g., in terms of features, rather than investigating the impact of sophisticated learning algorithms.

Recently, BERT, a pre-trained deep neural network (Devlin et al., 2019), based on the Transformer (Vaswani et al., 2017), has improved the state of the art for many natural language processing tasks. For example, it reached a score of 80.4 on the GLUE benchmark[1], 86.7% accuracy on MultiNLI, and F_1=93.2 on the SQuAD v1.1 question answering task. Currently, the top 11 systems in the SQuAD v2.0 use BERT.[2]

[1]https://gluebenchmark.com/.
[2]http://rajpurkar.github.io/SQuAD-explorer/.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1012–1015
Minneapolis, Minnesota, USA, June 6–7, 2019.

3 Method

We hypothesize that hyperpartisanship and extreme bias detection are related to sentiment analysis, which is one of the tasks in the GLUE benchmark. Given the recent success of BERT (Devlin et al., 2019) for sentiment analysis and other language processing tasks, we decided to experiment with it for hyperpartisan news detection.

In order to have a reference, we also experimented with Random Forests over TF.IDF representations. We used two BERT models: BERT without fine-tuning, and BERT with fine-tuning. We describe them in more detail below. In each case, we extracted features from the title and from the main text of the articles separately.

3.1 TF.IDF Features

In order to have a reference to compare our BERT-based approaches to, we also experimented with word-level TF.IDF features. First, we converted all text to lowercase and we stemmed it with the Porter stemmer. Then, we removed words with document frequency higher than 0.8. We then extracted two feature vectors by computing the term frequency and the inverse document frequency once on the title and separately on the body of the articles. We ended up with feature vectors of size 110,229 for the title and 1,798,179 for the content when TF.IDF vectors were computed on *BP-train* and *BA*, and 95,806 for the title and 1,507,789 for the content, when *BA* only was used.

We used the feature vectors in a Random Forest classifier with 100 estimators. Note that, differently from BERT, the TF.IDF representation is able to use information from the entire article.

3.2 Pre-trained BERT Features

Our second approach uses features extracted from Google's BERT, a model with pre-training language representations (Devlin et al., 2019). We fed to the model (*i*) the entire title and (*ii*) the first 256 tokens from the body of the article as two separate inputs, and then we obtained vector representations from the last layer of the BERT neural network. Note that we used the pre-trained BERT rather than training it with the data from the competition. Next, we concatenated the vectorial representations and we fed them to a two-layer feed-forward neural network with 32 neurons in the hidden layer. We used $\tanh$ as the activation function and a Gaussian noise with $\sigma = 0.2$.

3.3 Fine-tuned BERT Features

A natural extension of the approach in Section 3.2 is to fine-tune the BERT model on the datasets of the competition, i.e., on *BP+BA*. We performed fine-tuning on the same input used for computing the TF.IDF representations and we obtained two models, one from the titles only and one from the content of the articles only. As we did for the pre-trained model, we concatenated the internal vector representations from the last layer for both models and we passed them to the second neural network as in Section 3.2.

4 Experiments

We performed a number of experiments in order to select the best models to submit as official runs for the competition. The best model for the by-publisher dataset was selected on *BP-val* after training the models on *BP-train*. Since there was no validation set for the by-article dataset and *BA* was too small to be divided into training and validation sets, we trained our models on *BP* and we selected the best-performing one on *BA*. Table 1 shows the obtained accuracy values on *BP-val* (By-publisher) and *BA* (By-article) datasets. As we can observe, the performance of the TF.IDF model is behind those when using BERT, both with and without fine-tuning.

As a result, we opted for the two BERT models, trained on *BP+BA*, as our submissions for the competition. Table 2 shows the results on the hidden test sets.

Model	By publisher	By article
TF.IDF	56.13%	56.12%
BERT (no tuning)	61.20%	60.93%
BERT (fine-tuned)	61.70%	61.30%

Table 1: **Validation results:** Accuracy for our TF.IDF and BERT models on the by-publisher validation (*BP-val*) and on the by-article (*BA*) sets. The training was performed on *BP-train* and *BP*, respectively.

Model	By publisher	By article
BERT (no tuning)	63.25%	64.49%
BERT (fine-tuned)	64.60%	64.50%

Table 2: **Testing results:** Accuracy on the hidden test sets for the BERT models we submitted. Both models were trained on *BP+BA*.

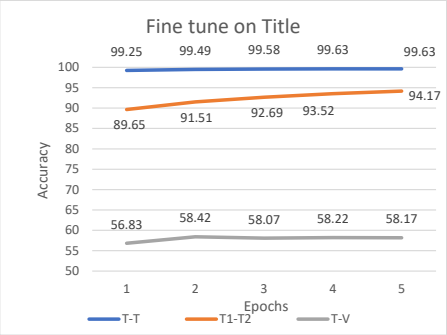

Figure 1: **Title as input:** Accuracy, at each epoch, for the fine-tuned BERT model. T-T stands for training and evaluating on *BP-train*, T1, T2 for training on the first half of *BP-train* and evaluating on the second half, T-V for training on *BP-train* and evaluating on *BP-val*.

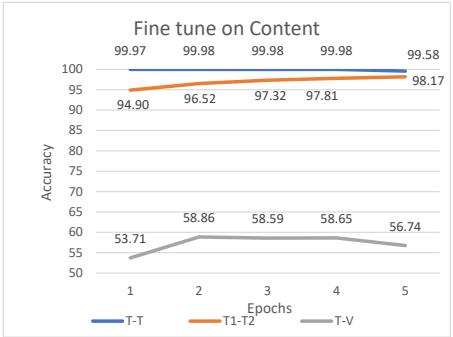

Figure 2: **Content as input:** Accuracy, at each epoch, for the fine-tuned BERT model. T-T stands for training and evaluating on *BP-train*, T1, T2 for training on the first half of *BP-train* and evaluating on the second half, T-V for training on *BP-train* and evaluating on *BP-val*.

4.1 Result Analysis and Post-Submission Experiments

When developing the model for the submission, we focused on the datasets with by-publisher annotation. This is probably the reason why we performed much better on the by-publisher hidden test set, 7th out 28 teams, than on the hidden by-article test set, 29th out of 42 teams.

Another possible reason for the low results on the by-article hidden test set is overfitting on *BP*: our model might have learned to discriminate the publishers appearing in *BP* instead of the required labels *hyperpartisan* / *non-hyperpartisan*. Recalling that *BP-val* does not contain any articles from the publishers in *BP-train*, we conducted an experiment to see whether there was a correlation between the articles in the different partitions of the provided dataset. In particular, we created a recurrent model with a single layer of 1,024 GRUs, and we trained it on 80% of the data and we evaluated it on the remaining 20%. The model achieved 99.99% accuracy at predicting whether the article was from *BP-train* or from *BP-val*.

We further performed three additional experiments with the fine-tuned BERT model: (*i*) training and evaluating on *BP-train*, (*ii*) training on the first half of *BP-train* and evaluating on the second half of *BP-train*, and (*iii*) training on *BP-train* and evaluating on *BP-val*.

Figure 1 shows the accuracy for BERT at each epoch when using titles for the three configurations (*i*)–(*iii*) above: T is *BP-train*, T1 and T2 are the two halves of *BP-train*, and V is *BA*.

Figure 2 reports the performance for the same experiments when using the body text of the articles as input. While the accuracy for the curves T-T and T1-T2 is close to 100% or is monotonically increasing with respect to the number of training epochs, the curve T-V does not show the same behavior, suggesting that 58.42% is close to the best performance that can be achieved in this setting. The accuracy values, although not directly comparable, show that there is a huge gap between the performance on a dataset with the same set of publishers (T-T and T1-T2) vs. on a dataset where the news comes from a different set of publishers (T-V), thus supporting our hypothesis.

5 Conclusions and Future Work

We have described our participation in SemEval-2019 task 4 on hyperpartisan news detection. In particular, we explored using TF.IDF and BERT-derived representations, and we found the latter to be more informative. Thus, we submitted two BERT models as our official runs to the competition: one with and one without fine-tuning. Interestingly, fine-tuning the model did not yield any sizable improvements. Our analysis suggests that our BERT models might be learning the source of the article, rather than whether it represents a piece of hyperpartisan news.

In future work, we plan to experiment with the big cased BERT model and to combine it with stylistic features, which have been proven successful for the hyperpartisanship detection task.

References

Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM*, 61(6):54–61.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 3528–3539, Brussels, Belgium.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, Minneapolis, MN, USA.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019a. Proppy: A system to unmask propaganda in online news. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI'19, Honolulu, HI, USA.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019b. Proppy: Organizing news coverage on the basis of their propagandistic content. *Information Processing and Management*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '19, Minneapolis, MN, USA.

Sean M. Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML '11, pages 489–496, Bellevue, Washington, USA.

Benjamin Horne, Sara Khedr, and Sibel Adali. 2018a. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM '18, pages 518–527, Stanford, CA, USA.

Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018b. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Proceedings of the The Web Conference*, WWW '18, pages 235–238, Lyon, France.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1113–1122, Baltimore, MD, USA.

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*, SemEval '19, Minneapolis, MN, USA.

Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 3518–3527, Brussels, Belgium.

Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On measuring bias in online information. *SIGMOD Rec.*, 46(4):16–21.

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*, ACL '18, pages 231–240, Melbourne, Australia.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 729–740, Vancouver, Canada.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 91–101, Seattle, WA, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.