Fine-Grained Propaganda Detection with Fine-Tuned BERT

Shehel Yoosuf,
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
syoosuf@mail.hbku.edu.qa

Yin "David" Yang
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
yyang@hbku.edu.qa

Abstract

This paper presents the winning solution of the Fragment Level Classification (FLC) task in the Fine Grained Propaganda Detection competition at the NLP4IF'19 workshop. The goal of the FLC task is to detect and classify textual segments that correspond to one of the 18 given propaganda techniques in a news articles dataset. The main idea of our solution is to perform word-level classification using fine-tuned BERT, a popular pre-trained language model. Besides presenting the model and its evaluation results, we also investigate the attention heads in the model, which provide insights into what the model learns, as well as aspects for potential improvements.[1]

1 Introduction

Propaganda is a type of informative communication with the goal of serving the interest of the information giver (i.e., the propagandist), and not necessarily the recipient (Jowett and O'donnell, 2018). Recently, Da San Martino et al. compiled a new dataset for training machine learning models, containing labeled instances of several common types of propaganda techniques. Through such fine-grained labels, the authors aim to alleviate the issue of noise arising from classifying at a coarse level, e.g., the whole article, as attempted in previous works on propaganda classification (Barrón-Cedeño et al., 2019; Rashkin et al., 2017). Using this dataset, the Fragment Level Classification (FLC) task of the Fine-Grained Propaganda Detection Challenge in NLP4IF'19 requires detecting and classifying textual fragments that correspond to at least one of the 18 given propaganda techniques (Da San Martino et al., 2019a). For instance, given the sentence "Manchin says Democrats acted like babies ...", the ground truth of FLC includes the detected propaganda technique for the fragment "babies", i.e., name-calling and labeling, as well as the start and end positions in the given text, i.e., from the 34th to the 39th characters in the sentence.

This paper describes the solution by the team "newspeak", which achieved the highest evaluation scores on both the development and test datasets of the FLC task. Our solution utilizes BERT (Devlin et al., 2018), a Transformer (Vaswani et al., 2017) based language model relying on multiheaded attention, and fine-tunes it for the purpose of the FLC task. One benefit of using the transformer architecture is that it leads to a more explainable model, especially with the fine grained information available through the dataset. We take a step in this direction by exploring the internals of the fine-tuned BERT model. To do so, we adapt the methods used in (Clark et al., 2019) and (Michel et al., 2019). In particular, we explore the average attention head distribution entropy, head importance, impact of masking out layers, and study the attention maps. The results reveal that the attention heads capture interpretable patterns, similar to ones observed in (Clark et al., 2019).

The rest of the paper is organized as follows. Section 2 presents our solution and elaborates on the architecture, training considerations and implementation details. Section 3 provides the results and analysis. Section 4 concludes with future directions.

2 Method

2.1 Solution Overview

We approach the problem by classifying each token in the input article into 20 token types, i.e., one for each of the 18 propaganda techniques,

[1] Code for reproducing the results can be found at https://github.com/shehel/BERT_propaganda_detection

Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 87–91
Hong Kong, China, November 4, 2019

a "background" token type that signifies that the corresponding token is not part of a propaganda technique, and another "auxiliary" type to handle WordPiece tokenization (Devlin et al., 2018). For example, the word "Federalist" is converted to "Federal" and "ist" tokens after tokenization, and the latter would be assigned the auxiliary token type. Since the labels provided in the dataset are at character level, before training our classifier, we first perform a pre-processing step that converts these character level labels to token level, which is later reversed during post-processing to obtain the outputs at the character level. This is done by keeping track of character indices of every word in the sentence.

The token classifier is obtained by fine-tuning a pre-trained BERT model with the input dataset and the token-level labels from the pre-processing step. Specifically, we add a linear classification head to the last layer of BERT for token classification. To limit training costs, we split articles by sentence and process each sentence independently in the subsequent token classifier. The classification results are combined in the post-processing step to obtain the final predictions, as mentioned earlier.

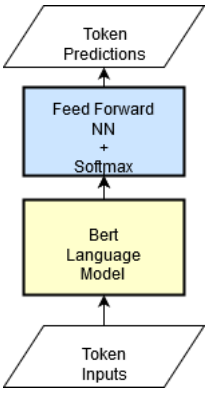


Figure 1: Architecture of our solution

2.2 Modeling

During the competition, we mainly explored three model architectures. The first is a simple scheme of fine-tuning a pre-trained BERT language model with a linear multilabel classification layer, as shown in Figure 1. The second performs unsupervised fine-tuning of the language model on the 1M news dataset (Corney et al., 2016) before supervised training on the competition dataset. This is motivated by the consideration of accounting for domain shift factors, since the BERT base model used in our solution was pretrained on BookCorpus and Wikipedia datasets (Devlin et al., 2018), whereas the dataset in this competition are news articles (Rietzler et al., 2019; Peters et al., 2019). Finally, the third model uses a single language model with 18 linear binary classification layers, one for each class. This is mainly to overcome the issue of overlapping labels, which is ignored in the former two model designs. Our final submission is based on the first architecture. Additionally, a fine-tuned BERT model with default parameters, i.e., the same setup described in the implementation section except for the learning rate schedule and sampling strategy, is used as a baseline for comparison in our experiments.

Preprocessing. Our solution performs token-level classification, while the data labels are at the character level. In our experiments, we observe that the conversion from character-level to token-level labels (for model fitting), as well as the reverse process (for prediction) incur a small performance penalty due to information lost in the conversion processes. Our final model in this competition does not consider overlapping labels, which occurs when one token belongs to multiple propaganda techniques simultaneously. Through experiments, we found that due to the above issues, the ground truth labels in the training data lead to an imperfect F1 score of 0.89 on the same dataset. This suggests that there is still much space for further improvement.

Dealing with Class Imbalance. The dataset provided in this competition is unbalanced with respect to propaganda classes. Some classes, such as "Strawmen", only have a few tens of training samples. To alleviate this problem, our solution employs two oversampling strategies: (i) weighting rarer classes with higher probability and (ii) sample propaganda sentences with a higher probability (say, 50% higher) than non-propaganda sentences. Such oversampling, however, also have adverse effects such as loss of precision and overfitting. Hence, the sampling method in our final submission strikes a balance through curriculum learning (Bengio et al., 2009), whereby an oversampling strategy is used in the first half of the training and sequential sampling is used in the second half.

Implementation. We trained all models on a machine with 4 Nvidia RTX 2080 Ti graphic cards. Our implementation is based on the Py-

Torch framework, using the pytorch-transformers package.[2] To accelerate training, all models were trained in mixed precision.

Our best models are based on the uncased base model of BERT which was found to work better than cased model, containing 12 transformer layers and 110 million parameters trained using the following hyper-parameters: batch size 64, sequence length 210, weight decay 0.01, and early stopping on F1 score on the validation set with patience value 7. Each model, including the final submission, was trained for 20 epochs. We used the Adam optimizer with a learning rate of 3e-5 and cosine annealing cycles with hard restarts and warmup of 10% of the total steps.

During the event, participants had only access to the training set labels which was split into a training set and a validation set with 30% of the articles chosen randomly. Models for submitting to the development set was chosen based on validation F1 scores, which in turn, informed the submissions for the test set.

2.3 Attention Analysis

We first measure the general change in behavior of the attention heads after finetuning on the dataset. We do this by visualizing the average entropy of each head's attention distribution before and after finetuning on the dataset. Intuitively, this measures how focused the attention weights of each of the heads are.

Next, we calculate head importance using

$$I_h = \mathbb{E}_{x \sim X} \left| \frac{\partial \mathcal{L}(x)}{\partial \xi_h} \right|, \quad (1)$$

where ξ_h is a binary mask applied to the multihead attention function to nullify it. X is the data distribution and $\mathcal{L}(x)$ is the loss on sample x. If I_h has a large value, it can be interpreted as an important head since changing it could also have a greater impact on the performance of the model. We use these scores to determine heads to visualize.

3 Results

The model that performed the best empirically was the BERT language model with a simple classifier, with parameter tuning, masked logits, cyclical learning rates and a sampling strategy. Table 1 shows the scores on the development set of the models we tried including the baseline BERT model. Retraining language model on 1M News dataset failed to match the performance of the original model. The model design with multiple binary classification linear layers (which is capable of predicting multiple labels for a token) obtained better results on some rarer classes; however, its performance on more common classes is lower, leading to a lower overall F1 score. However, we cannot draw conclusions on these approaches as we hypothesize that this could be improved by using a more optimal learning approach.

Model	F1	Precision	Recall
BERT-baseline	0.2214	0.252	0.1972
BERT-18 Binary	0.2273	0.2603	0.2017
BERT-1M News	0.2078	0.2671	0.17
BERT-submission	**0.242**	**0.289**	**0.208**

Table 1: Evaluation results on official development set

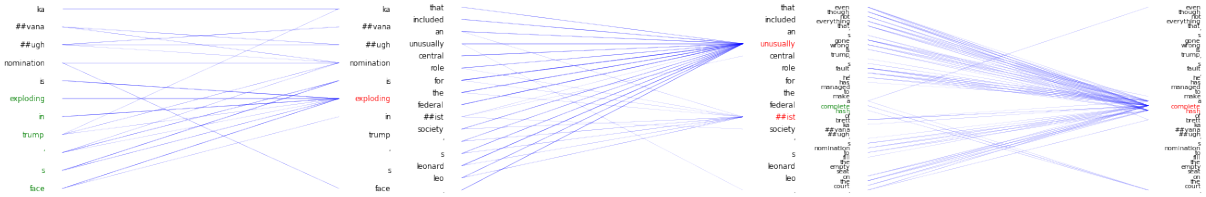
Technique	Dev F1	Test F1
Appeal-Authority	0	0
Appeal-Fear	0.3268	0.209
Bandwagon	0	0
Black-White-Fallacy	0	0.09
Casual-Oversimplification	0.05	0
Doubt	0.125	0.169
Exaggeration-Minimisation	0.276	0.159
Flag-Waving	0.409	0.438
Loaded-Language	0.4078	0.331
Namecalling-Labeling	0.2605	0.394
Obfuscation-Confusion	0	0
Red-Herring	0	0
Reductio-Hitlerum	0.206	0
Repetition	0.014	0.011
Slogans	0.153	0.1305
Strawmen	0	0
Thought-Cliches	0	0
Whataboutism	0.16	0

Table 2: Classwise F1 scores for final submission

The model with the highest score based on BERT with a single multilabel token classification head was chosen as our submission to evaluate on the test set which yielded a test F1 score of 0.2488, 0.286 precision and 0.22 recall (see table 2 for class wise scores). This model won the competition.

We improved on the strong performance of baseline BERT model by firstly using an oversampling strategy where sentences with propaganda are weighted more, which in our final submission was 50%. Such an approach works because the number of sentences with no propaganda is much higher than that of ones with propaganda. Attempts at fixing the imbalance among propaganda techniques was found to be detrimental for the purpose of this competition, because the evaluation metric does not take into account this imbalance. Although oversampling helped the model

[2] https://github.com/huggingface/pytorch-transformers



Layer 9 Head 8 Layer 11 Head 4 Layer 11 Head 4

Figure 2: Attention maps labeled by their layer and head number respectively. Green highlights propaganda fragment and red highlights the behaviour. The darkness of the line corresponds to the strength of the weight.

learn, we found that this led to overfitting and the model losing precision. Ablation studies showed that following oversampling with sequential sampling did indeed help improve the precision of the system. Second, we used an appropriate cyclic learning rate scheme to avoid poor local minima (Smith, 2017) as explained in previous section.

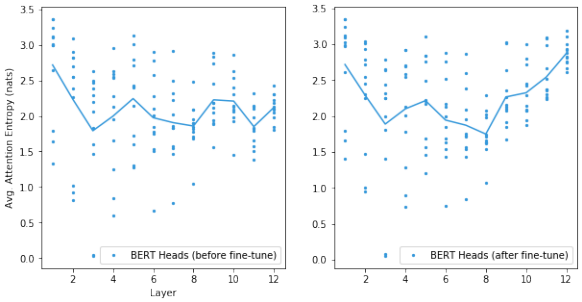


Figure 3: Average entropy of the attention weights of every attention head across layers

We examined attention heads from different layers based on their importance score. Excluding the linguistic patterns reported in (Clark et al., 2019), additional task specific patterns were observed indicating the model's ability to represent complex relationships (See Fig 2). For example, a number of heads appear to attend to adjectives and adverbs that could be useful for several propaganda techniques. Similarly, some heads pick out certain "loaded" words which all words in the sentence strongly attend to. However, it should be noted that the roles of attention heads are not clear cut, and further experimentation is required to further study this issue.

Next, we calculated the average entropy of the attention distribution of heads before and after fine-tuning. Fig 3 shows the entropy after the 8th layer increasing after fine-tuning while the earlier layers remain almost unchanged. It also happens that most of the high importance ranked heads are clustered between layers 5 and 8. We tried masking out the last 4 layers and fine-tuning the model giving an F1 score of 0.2 on the development set. This leads us to believe that BERT is still undertrained after fine-tuning as explored in (Liu et al., 2019) and requires better training strategies and hyperparameter selection schemes to fully utilize it.

4 Conclusion and Future Work

This paper describe our winning solution in the Fragment Level Classification (FLC) task of the Fine-Grained Propaganda Detection Challenge in NLP4IF'19. Our approach is based on the BERT language model, which exhibits strong performance out of the box. We explored several techniques and architectures to improve on the baseline, and performed attention analysis methods to explore the model. Our work highlights the difficulty of applying overparameterized models which can easily lead to sub-optimal utilization as shown in our analysis. The results confirm that language models are clearly a step forward for NLP in terms of linguistic modeling evident from its strong performance in detecting complex propaganda techniques.

Regarding future work, we plan to explore further methods for parameter efficient modeling which we hypothesize as being key for capturing interpretable linguistic patterns and consequently better representations. One related direction of research is spanBERT (Joshi et al., 2019), which includes a pretraining phase consisting of predicting spans instead of tokens which is inherently more suited for the propaganda dataset. Further, we plan to investigate methods and models that are capable of capturing features across multiple sentences, which are important for detecting some propaganda classes such as repetition. Finally, we also plan to look into visualizing and identifying

additional patterns from the attention heads.

Acknowledgments

This publication was made possible by NPRP grant NPRP10-0208-170408 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work, and are solely the responsibility, of the authors.

References

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

David Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. 2016. What do a million news articles look like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, pages 42–47.

Giovanni Da San Martino, Alberto Barron-Cedeno, and Preslav Nakov. 2019a. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom (NLP4IF): Censorship, Disinformation, and Propaganda*, NLP4IFEMNLP '19, Hong Kong, China.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019, Hong Kong, China.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Garth S Jowett and Victoria O'donnell. 2018. *Propaganda & persuasion*. Sage Publications.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *arXiv preprint arXiv:1905.10650*.

Matthew Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

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*, pages 2931–2937.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*.

Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.