TTUI at SemEval-2020 Task 11: Propaganda Detection with Transfer learning and Ensembles

Moonsung Kim and **Steven Bethard**
School of Information
College of Social & Behavioral Sciences
University of Arizona
{moonsungkim,bethard}@arizona.edu

Abstract

In this paper, we describe our approaches and systems for the SemEval-2020 Task 11 on propaganda technique detection. We fine-tuned BERT and RoBERTa pre-trained models then merged them with an average ensemble. We conducted several experiments for input representations dealing with long texts and preserving context as well as for the imbalanced class problem. Our system ranked 20th out of 36 teams with 0.398 F1 in the SI task and 14th out of 31 teams with 0.556 F1 in the TC task. Our code is available at https://github.com/amenra99/SemEval2020_Task11.

1 Introduction

Propaganda is purposeful information aiming to influence an audience to persuade an ideological or political agenda. Propagandistic messages use psychological and rhetorical techniques to hide their intention and can be delivered through materials such as articles, books, images, videos, etc. There have been multiple studies addressing propaganda detection using machine learning systems (Rashkin et al., 2017; Volkova and Jang, 2018; Barrón-Cedeno et al., 2019; Da San Martino et al., 2019b). More recently, Da San Martino et al. (2019a) organized the NLP4IF-2019 shared task on Fine-Grained Propaganda Detection extended from their previous work, which has a similar framework to this shared task.

In SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, the organizers provided a Propaganda Techniques Corpus (PTC) manually annotated by six professional annotators in 446 news articles from 48 news outlets (Da San Martino et al., 2019b). This shared task is divided into two sub-tasks. The first sub-task is Span Identification (SI) to identify the begin offset and end offset of propaganda fragments in text. This binary sequence tagging task is evaluated by the F1 score of the normalized length of the intersection between manually annotated and predicted spans. The second one is Technique Classification (TC) to label the name of the technique of a given propaganda fragment. In this sub-task, a system is required to classify the specific propaganda fragment into one of the 14 given classes. The original classes were 18 but the organizers merged five underrepresented techniques into two classes to resolve a low proportion of some techniques. This multi-class classification task is evaluated based on the micro-average F1 score. Both sub-tasks share the same dataset and only the last submission on the test set is evaluated per sub-task even though there are no limitations of the number of submissions. A more detailed description of this shared task can be found in Da San Martino et al. (2020).

In this paper, we describe our approaches toward dealing with these problems. Two state-of-the-art transformer-based language models, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and a Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019), were used in this shared task. We fine-tuned multiple pre-trained models and merged the models using an average ensemble. We also conducted some experiments to find out which input representations preserve context better. Our final performance on the test set ranked in 20th out of 36 with 0.398 F1-score for the SI task and in 14th out of 31 with 0.556 for the TC task.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1829–1834
Barcelona, Spain (Online), December 12, 2020.

2 Methodology

2.1 Fine Tuned Models

In both subtasks, we utilized the transformer-based models BERT and RoBERTa as they have achieved successful performance on several natural language processing tasks recently. Because both models released multiple versions of pre-trained models trained on the different corpora, we examined all the versions to find the proper one for the dataset; The uncased version of BERT-base and BERT-large and RoBERTa-base and RoBERTa-large models were tested in this shared task. These models were fine-tuned on the training set with a learning rate of 2e-5 for 10 epochs. Then, we took the best epoch per model based on the F1 score on the development set. All models were trained on Nvidia P100 using the Tensorflow version of the Huggingface Transformers library (Wolf et al., 2019).

After that, we applied an average ensemble on the results across the models. Empirically, bagging ensemble techniques allow better performance and minimize over-fitting (Opitz and Maclin, 1999). Most participants with top-ranked performance in many other machine learning competitions used this technique. We averaged the probability distributions of the classifiers and then took the maximum prediction over the average probability distribution. We adopt the best-performing ensemble models per each task on the development set for our final models on the test set.

2.2 Span Identification (SI)

Data encoding: The dataset was encoded using tokenizers provided by the BERT and RoBERTa pre-trained models. The tokenizers break strings into sub-word tokens. Then we tagged each token with a binary label according to whether it was a part of a propaganda fragment or not. The span of each token was recorded to keep track of the position for prediction. An example input for the SI task looks like:

Sentence	Different	rules	for	the	**lawless**		**party**	.
Token	different	rules	for	the	law	less	party	.
Label	0	0	0	0	1	1	1	0
Span	(0, 9)	(10, 15)	(16, 19)	(20, 23)	(24, 27)	(27, 31)	(32, 37)	(37, 38)

Input segmentation: Dealing with long texts was one of the challenges in this binary sequence tagging task. While the maximum sequence length of BERT and RoBERTa is 512, the token length of each news article exceeds this. We need to split the dataset into smaller chunks to feed into the models. Whether the token is propaganda is determined by its context in the text and it is more context-sensitive than Part-of-Speech (POS) tagging or Named Entity Recognition (NER) tasks since it is sparser and has less consistency across different instances of a word. To keep contextual information in this sub-task, we applied the overlapped token feeding method as this technique on BERT models outperformed other approaches in long sequence classification tasks (Pappagari et al., 2019). Sentence-level overlapping was used rather than fixed token length since each article from the dataset was provided with split sentences. We conducted an experiment to find out the optimal number of sentences to overlap. The methods we tried are as follows (illustrated in fig. 1).

Sentence by Sentence feed one sentence at a time.

Sentence Chunk split text into sentences chunks in which the total number of the tokens less than 512 without overlapping.

Overlapped Sentence Chunk Same as Sentence Chunk but overlap with the last n sentences from the previous chunk in the same article ($1 \leq n \leq 4$). The OR operation is taken on the prediction of each token of overlapped sentences to resolve conflicts.

We recorded the F1 score of each method using the uncased BERT-base model at the 5th epoch on the development set with a max sentence length of 512 and a batch size of 8. The overall results on the development set are shown in table 1. Our observation from this experiment is that the overlapped feeding method is preferred in the sequence labeling task. We applied chunking with 4 previous sentences overlapping in our model development since it showed slightly better performance than others.

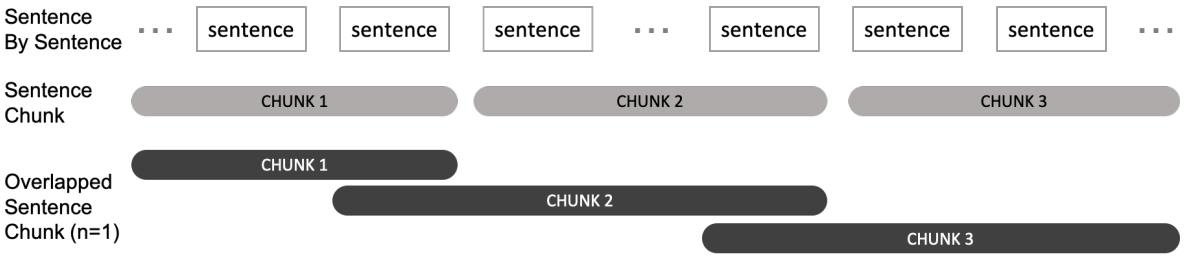

Figure 1: Sentence segmentation methods

Post processing: After reviewing our model predictions on the development set with the gold labels, we realized our model failed to predict some intervening words. For example, in a gold fragment "preaching of hatred and jihad violence", our model missed tagging "and" as a propaganda word. Most of these failures were prepositions or conjunctions like "of", "to", "on", "for", etc. Since the evaluation of this sub-task measured the length of intersection between the gold and the prediction spans, we applied a heuristic to label the word as propaganda if the tokens immediately before and after the target word were a part of propaganda fragments. This technique provided some minor improvement in F1 score on the development set (see table 2).

2.3 Technique Classification (TC)

The goal of this sequence classification sub-task is to discriminate the type of propaganda. We used the same tokenizers as the SI task for encoding the dataset, but each input was assigned one of the given 14 classes. The class labels are listed in table 6.

Context Feature: As in the SI task, we examined the effectiveness of applying context features in this sequence classification task. The experiment was conducted with the following three conditions:

Without Context feed only the target fragment without context sentence. In the example text below, only the *"where free thought goes to die"* part is used as an input.

> *. . . Combine it with billions of Facebook and Twitter users and you can easily see the influence wielded here. This is* ***where free thought goes to die****. Ron Paul reminds us that truth is treason. . . .*

Parent Sentence include the sentence that contains the target fragment. The parent sentence is concatenated after the target fragment following BERT's separation token [SEP]. For instance, the input looks like: *[CLS]where free thought goes to die[SEP]This is where free thought goes to die.[SEP]*

Method	F1
Sentence by sentence	0.368
Sentence chunk	0.390
1 sentence overlap	0.393
2 sentence overlap	0.399
3 sentence overlap	0.405
4 sentence overlap	0.410

Table 1: Input segmentation for Task SI on development set

Method	Example	F1
Original prediction	***smear*** *and* ***assassinate***	0.449
Filling intervening words	***smear and assassinate***	0.454

Table 2: Post processing for Task SI on development set

Method	F1
Only fragment	0.534
Fragment + Parent sentence	0.525
Fragment + Before sent + Parent sent + After sent	0.506

Table 3: Context input for Task TC on development set

Models	F1-score	
	SI	TC
BERT-base-uncased	0.392	0.553
BERT-large-uncased	0.449	0.568
RoBERTa-base	0.429	0.560
RoBERTa-large	0.476	0.305
BERT-base-uncased + RoBERTa-base	0.424	0.575
BERT-base-uncased + RoBERTa-large	0.490	0.525
BERT-large-uncased + RoBERTa-base	0.472	0.581
BERT-large-uncased + RoBERTa-large	0.496	0.506

Table 4: Fine-tuned model performance on the development set

Before and After Sentences include one previous and one following sentence along with the parent sentence of the target fragment. In this condition, the input is *[CLS]where free thought goes to die[SEP]This is where free thought goes to die.[SEP]Combine it with billions of Facebook and Twitter users and you can easily see the influence wielded here.[SEP]Ron Paul reminds us that truth is treason.[SEP].*

Table 3 shows the results at the 5th epoch on the uncased BERT-base model. We found that adding context features provided no significant improvements at this time. Rather, the prediction performance was decreased with the surrounding contextual sentence concatenation. This failing may result from multiple propaganda fragments being contained in the same sentence - or in the previous or the following sentence, which makes it difficult to capture a signal; 316 fragments out of 6,129 are located in the same sentence, and almost 80 percent of fragments have other fragments in the surrounding sentences in the training dataset. Thus, only the fragment without context was used as an input in our final model.

Class Imbalance: The imbalance of classes was an issue in this sub-task. The top five propaganda classes accounted for 78 percent of the total, whereas the rest were each under 5 percent. To address this problem, we applied class weights by *w = number_of_samples / (number_of_occurrences_per_class * number_of_classes)* to assign higher weights on infrequent classes and lower weights on more representative classes. This gave some improvement in the overall F1 score on the development set (see section 3.2).

3 Results and Discussion

3.1 SI Results

We trained the final SI model using the best hyperparameters from the experiments on the development set, but trained on the combined training and development data. Based on the model performance comparison (table 4), the ensemble of uncased BERT-large and RoBERTa-large model was used on both datasets. Table 5 shows the performance of our SI models on the development and the test dataset. There was a significant performance drop between the development set and the test set. The final result on the test set was F1 score of 0.398, which ranked 20th out of 36 teams. Our model achieved the highest precision among all participants, but with poor recall. The results on the development set showed a better

Data	Our system			Top system		
	F1	Precision	Recall	F1	Precision	Recall
Development set	0.513	0.495	0.533	0.534	0.399	0.808
Test set	0.398	0.669	0.284	0.516	0.565	0.474

Table 5: Task SI results on development set and test set

Propaganda Techniques	Distribution		Development		Test	
	Freq.	Ratio	No Weights	Class Weights	Our System	Top System
Appeal_to_Authority	144	2%	0.077	0.162	0.282	**0.481**
Appeal_to_fear-prejudice	294	5%	0.269	0.347	0.415	**0.455**
Bandwagon,Reductio_ad_hitlerum	72	1%	0.000	0.667	**0.246**	0.083
Black-and-White_Fallacy	107	2%	0.063	0.190	0.353	**0.490**
Causal_Oversimplification	209	3%	0.400	0.333	**0.231**	0.227
Doubt	493	8%	0.497	0.569	**0.574**	0.562
Exaggeration,Minimisation	466	8%	0.430	0.420	0.322	**0.336**
Flag-Waving	229	4%	0.798	0.775	0.617	**0.694**
Loaded_Language	2123	35%	0.737	0.751	0.732	**0.771**
Name_Calling,Labeling	1058	17%	0.688	0.710	0.685	**0.744**
Repetition	621	10%	0.287	0.321	0.212	**0.545**
Slogans	129	2%	0.320	0.500	0.375	**0.513**
Thought-terminating_Cliches	76	1%	0.343	0.207	0.250	**0.392**
Whataboutism,Straw_Men,Red_Herring	108	2%	0.105	0.291	0.203	**0.250**
Overall	6129	100%	0.575	0.590	0.556	**0.621**

Table 6: Propaganda class distribution and F1 scores on development and test set for Task TC

balance of precision and recall. Further analysis is required to understand why training on more data (training+development) with the same parameters works so differently on the test data.

3.2 TC Results

We trained the final TC model using the best hyperparameters from the experiments on the development set, but trained on just the training data. As shown in table 4, the uncased BERT-large and RoBERTa-base ensemble performed the best for the TC task. The RoBERTa-large model showed poor prediction capability in the TC task, unlike in the SI task. Our model achieved a micro-averaged F1 score of 0.556 and ranked 14th out of 31 teams overall on the test set (see table 6). One observation from the leaderboard was that our model successfully predicted at least some instances of each class whereas some high ranked teams failed to predict a few minority classes. However, classes with minority proportions still had low scores even though class weights were applied to deal with the imbalanced class problem. Future work should investigate how to increase performance on a dataset that has an unbalanced distribution.

4 Conclusion

We presented a description of our systems to identify propaganda fragments in news articles. The average ensemble models of BERT and RoBERTa performed better than when they were used independently in both sub-tasks. To deal with long articles in these models, we observed chunking sentences with overlapping 4 previous sentences showed the best performance among other segmentation methods.

As a conclusion, our work showed competitive results in this shared task but still has limitations. Training with the development dataset for the testing model should be treated carefully due to its difficulty in selecting an effective model. And predicting minority classes was another challenge in achieving high performance.

5 Acknowledgements

This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII) and maintained by the UArizona Research Technologies department.

References

Alberto Barrón-Cedeno, 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.

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 Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170.

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, November.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval 2020, Barcelona, Spain, September.

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*.

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*.

David Opitz and Richard Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198.

Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. *arXiv preprint arXiv:1910.10781*.

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.

Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*.