Inno at SemEval-2020 Task 11: Leveraging Pure Transfomer for Multi-Class Propaganda Detection

Dmitry Grigorev
Innopolis University
d.grigorev@innopolis.ru

Vladimir Ivanov
Innopolis University
v.ivanov@innopolis.ru

Abstract

The paper presents the solution of team ”Inno” to a SEMEVAL 2020 task 11 ”Detection of propaganda techniques in news articles”. The goal of the second subtask is to classify textual segments that correspond to one of the 18 given propaganda techniques in news articles dataset. We tested a pure Transformer-based model with an optimized learning scheme on the ability to distinguish propaganda techniques between each other. Our model showed 0.6 and 0.58 overall F1 score on validation set and test set accordingly and non-zero F1 score on each class on both sets.

1 Introduction

Modern society is experiencing an overflow of different kinds of information. Thousands of news media are publishing a giant number of articles, blogposts, twits and other kind of content every day and even every hour. People who want to stay informed about global situation are often not able to distinguish which source is reliable and provides relevant content. Therefore, such topics as automatic fake news detection, sentiment analysis and propaganda detection are gaining increasing attention.

The authors of (Jowett and O'donnell, 2018) defined propaganda in the most neutral sense as dissemination or promotion of a particular idea.

This paper provides our solution to SEMEVAL 2020 task 11 “Detection of propaganda techniques in news articles”. The task consists of two subtasks: (i) a propaganda sample span identification and bounding and (ii) a classification of these spans into one of the 18 classes (or techniques). Some techniques in the second subtask were merged into one class to get 14 classes for prediction in the end. All the techniques, as well as the other details about the task, are described by the task organizers in (Da San Martino et al., 2020).

In this paper we present a solution to the second subtask. The suggested system[1] consists of a Transformer-based classifier with additional learning optimization techniques such as undersampling, cost-sensitive learning and context addition, aimed at robustness. This approach allowed us to achieve 0.6 and 0.58 overall F1 scores with non-zero F1 score for each class on validation set and test set accordingly. We took the 7th place out of 31 teams on the second subtask.

2 Background

One of the first classifications of propaganda techniques was proposed in 1936 by Clyde R. Miller (Edwards, 1938) after a presidential election in the USA. This categorization consists of description of 7 propaganda techniques: “Name calling”, “Glittering generalities”, “Transfer”, “Testimonial”, “Plain folks”, “Card stacking” and “Band wagon”. Since then a lot more techniques have been identified, but this classification is still relevant (some of the techniques have been included in the SEMEVAL 2020). Probably, the most promising state-of-art strategies of dealing with propaganda and filtering it from sentimentally neutral texts is the one of viewing the task as a text classification problem. Some of the research efforts in propagandist texts classification are described further in this section.

[1] Source code can be found here https://github.com/oxxford/SEMEVAL-2020

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1481–1487
Barcelona, Spain (Online), December 12, 2020.

Approach	Times it was used
BERT	13
LSTM	7
Language features	5
ELMo	3
Logistic regression	3

Table 1: How often different approaches were used

(Rashkin et al., 2017) investigated news sources of four categories - truthful, propagandist, hoax and satiric - to show what language features correspond to each news source category. They examined 50 highest weighted n-gram features in the MaxEntropy classifier for each class and showed that truthful sources heavily use specific places (e.g. "London") as well as exact time (e.g. "on Tuesday"). For propagandist news features the sources employ abstract generalities (e.g. "truth", "freedom") and specific issues (e.g. "election", "shooting"), while for satiric articles vaguely facetious hearsay (e. g. "reportedly", "confirmed") are employed. Finally, hoax publishers utilize divisive topics (e.g. "liberals", "trump") and dramatic cues (e.g. "breaking"). The authors also proposed that hoax and propagandist sources often use videos as a reference since words like "video" and "youtube" are also included in the highest weighted for these categories.

(Barrón-Cedeno et al., 2019) have significantly contributed to the field. They not only collected a novel dataset that overall contains more than 35000 news articles both with and without propagandist content, but also compared what text features allow to efficiently classify articles on propagandist and non-propagandist. As a result they concluded that stylistic features like character n-grams show better results than topic-related features. The authors also demonstrated a service for news articles propaganda level assessment.

(Martino et al., 2019) have proposed another corpus for propaganda classification. It consists of 451 articles from 48 sources. Every article was manually annotated sentence by sentence by identifying the presence or absence of one of 18 propaganda techniques (grouped into 14 classes). The authors also presented baseline models for binary and multi-class classification based on BERT (Devlin et al., 2018) and launched a task on propaganda detection that will be described later.

A major step in the field of propaganda detection was an establishment of Fine-Grained Propaganda Detection challenge (Da San Martino et al., 2019). This challenge, similar to the current one, had two tasks:

- The first one was to classify sentences into propaganda and non-propaganda classes. The sentences were taken from the articles from propagandist and non-propagandist sources. The metric was F1 score.
- The second task consisted of two parts - extract spans that contain propaganda from article text and classify what propaganda technique is used. The metric took into account both span borders and class predictions. In the current version of the challenge the task was presented by two separate tasks, span identification and classification.

Overall, 25 teams submitted their solutions of the first task and 12 teams submitted their solutions of the second task. Leading teams used Transformer-based (Vaswani et al., 2017) models that are built on a powerful encoder-decoder architecture having self-attention mechanism as its core. This mechanism allows to take into account inter-sequence word dependencies, providing rich semantic representation. Almost all teams used some Transformer-based models (especially BERT (Devlin et al., 2018)) either to get embeddings or as a pretrained model (Yoosuf and Yang, 2019) (Hou and Chen, 2019). Other teams often used ensembles with different features and models inside: LSTM-CRF (Gupta et al., 2019), XGBoost (Tayyar Madabushi et al., 2019), BiLSTM (Vlad et al., 2019) and others. Table 1 summarizes approaches extensively used in the solutions.

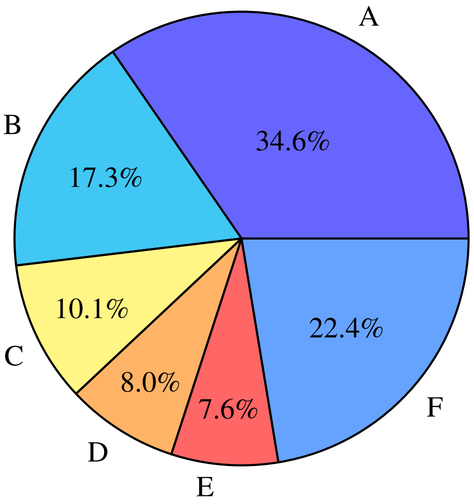

Figure 1: Class distribution in the train data, where A is "Loaded Language", B is "Name Calling or Labeling", C is "Repetition", D is "Doubt", E is "Exaggeration or Minimisation", and F represents all the remaining 9 classes.

As Table 1 shows, Transformer-based models (BERT in particular) have become a leading architecture in propaganda detection field.

3 System overview

Considering the top-performing systems of the previous challenge, we decided to use Transformer-based model as a solution baseline - mainly focusing on BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). However, since a lot of new models have been published since the previous challenge, we also tested several different Transformer-based models, including DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019), XLnet (Yang et al., 2019). After calculating F1 scores for all of them, we chose the one with the highest performance - RoBERTa. We also decided to use a model pretrained on a huge dataset and fine-tune it on the given dataset by adding a fully-connected classification layer. This strategy is highly suitable for our case because the training dataset consisted of 371 articles and contained 6129 propaganda samples in total for all classes, which is not enough for training a good-performing Transformer model from scratch, but is reasonable for fine-tuning. The following sections describe the techniques we used for optimizing the robustness of our model.

3.1 Imbalanced Data and Undersampling

One essential feature of the dataset is high class imbalance. Figure 1 demonstrates class distribution among samples.

As our first attempt to overcome class imbalance we used undersampling technique, which implies reducing the portion of more frequent class in the data by removing some samples. In our case, the most frequent classes were 'Loaded Language' (2123 out of 6129 samples in train set), and 'Name Calling,Labeling' (1058 out of 6129 samples in train set). Our aim was to retain approximately few hundred samples for each of these classes. To obtain this we experimented with different coefficients: 0.35, 0.3, 0.2 for 'Loaded Language' and 0.4, 0.5, 0.6 for 'Name Calling, Labeling'. The best combination with respect to final F1 score was 0.2 for 'Loaded Language' and 0.5 for 'Name Calling,Labeling'.

3.2 Cost-sensitive learning

Another technique we tried to deal with class imbalance was cost-sensitive learning. The idea of this method is to assign different "cost" to classes with different inclusion percentage in the dataset. We implemented the method by assigning label weights during loss calculation. The weight w_i for each label was calculated using formula (formula 3.2):

$$w_i = \frac{c_i}{\sum_j c_j}, c_i = \frac{\sum_j l_j}{l_i}$$

where l_i is number of samples of label i in the dataset. This formula (formula 3.2) can be interpreted like this: class weight is inversely proportional to the portion of this class in the dataset.

3.3 Context addition

Having performed the dataset analysis we found out that span samples may highly vary in length: from one-word long samples to several sentences long sentences. Moreover, this significant difference may occur within the same class. Figure 2 shows box-plots of sample lengths grouped by classes. It is interesting that three most frequent classes (8, 9, 10) have the shortest average lengths.

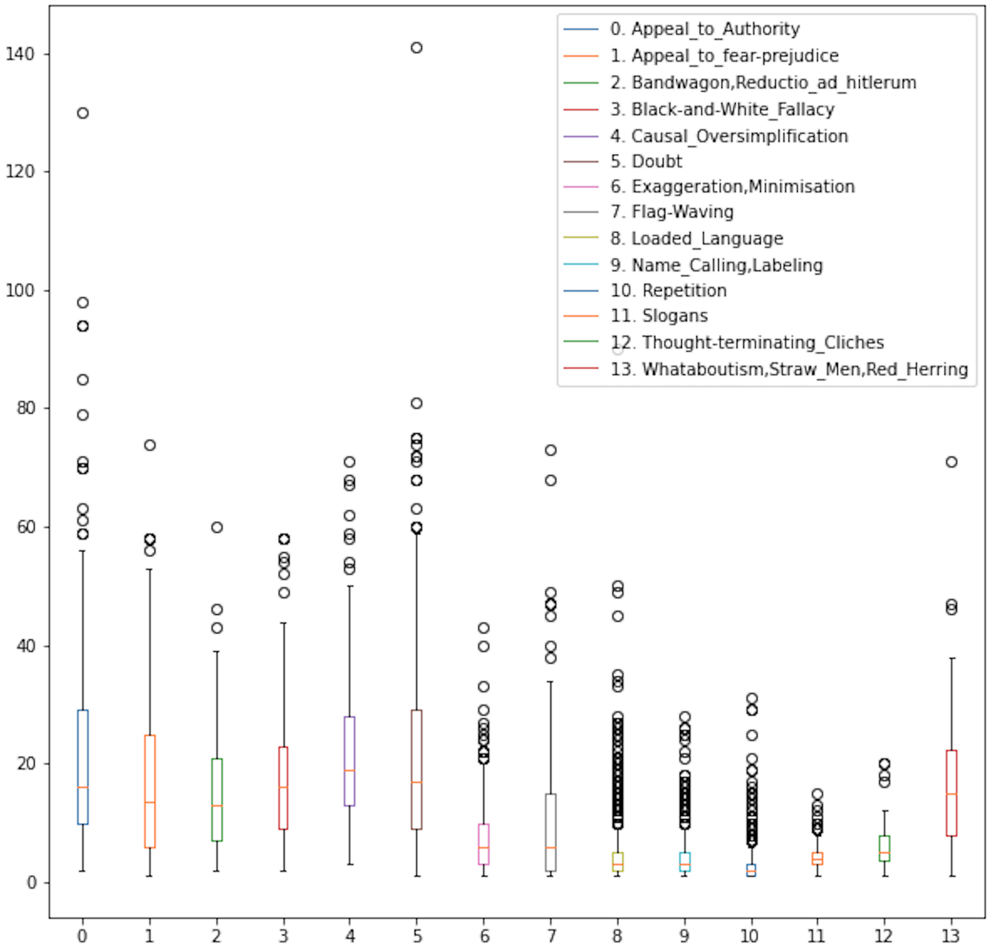

Figure 2: Box-plots of length distribution in words inside classes

Our expectation was that spans with such a diverse lengths cannot efficiently represent all techniques and that some useful information can be left outside the span boundaries. To check this, we decided to expand the spans with the surrounding context. We tried two ways of setting the new sample boundaries: to complete a sentence (any text surrounded by '.', '?', '!' signs or "end of line" symbol on both sides) and to complete a sentence or a subsentence (same as complete a sentence but with ',' sign included). However, this spans expansion strategy was not efficient as it introduced more noise and, thus, reduced the score.

4 Experimental Setup and Results

To choose between different Transformer-based models we trained each one of them on a single epoch and compared their validation scores. The result of this evaluation are shown in Table 2.

Model	F1 val. score
ALBERT	0.498
DistillBERT	0.502
XLnet	0.526
BERT	0.553
RoBERTa	**0.573**

Table 2: Baseline models performance

Strategy	F1 val. score
Baseline	0.57384
Tuned	0.59172
Tuned+Undersampling	0.59266
Tuned+Cost-sensitive learning	**0.60113**
Tuned+Context addition	0.55691

Table 3: Performance with different techniques applied

According to Table 2, RoBERTa showed the best performance results among all the other models, therefore we chose it as a model for future improvement. It is also worth mentioning that all models showed similar patterns on the same classes:

- High performance ($F_1 \geq 0.5$) on ”Flag-Waving”, ”Loaded Language” and ”Name Calling, Labeling”;

- Intermediate performance ($0.1 \leq F_1 < 0.5$) on ”Appeal to fear-prejudice”, ”Doubt”, ”Exaggeration, Minimisation” and ”Repetition”;

- Low performance ($F_1 < 0.1$) on ”Appeal to Authority”, ”Bandwagon, Reductio ad hitlerum”, ”Black-and-White Fallacy”, ”Causal Oversimplification”, ”Slogans”, ”Thought-terminating Cliches” and ”Whataboutism, Straw Men, Red Herring”.

As the next step, we tuned the hyperparameters of the model including the number of epochs, batch size and maximum sequence length. This step allowed us to improve F1 score for 2 points and provided us with a ready-to-run model so that we could proceed to optimization techniques described in previous section. Table 3 shows the result of applying this strategy.

4.1 Our submission

Our final system submitted for this task is based on RoBERTa model, combined with cost-sensitive learning strategy. We took pretrained weights from Transformers (Wolf et al., 2019) library by Hugging Face, and fine-tuned the model on 5 epochs with batch size 8 and maximum sequence length of 512. We used Simple Transformers[2] library for training and inference.

The submitted model showed 0.58 F1 score on the test set taking the 7th place among 31 teams in the second track. The system also showed non-zero F1 score on each class (some models with better final performance did not achieve this), which means that the model can capture some meaning of each propaganda technique. It is also worth mentioning that the difference between validation performance and test performance of our model is 2 points, which is less than that of most models with better test performance and, moreover, it is less than that of models in top three. Thus, we achieved less overfit compared to the competitors.

[2] https://github.com/ThilinaRajapakse/simpletransformers

5 Conclusion

In this paper we described a system that we developed to participate in SEMEVAL 2020 Task 11 “Detection of propaganda techniques in news articles”. We trained a pure Transformer-based model with preprocessing and achieved 0.58 F1 score on test data and non-zero F1 score for each class. The approach demonstrated the effectiveness of combining a Transformer-based baseline with learning optimization techniques including undersampling, cost-sensitive learning and context augmentation.

We also demonstrated that compared to other NLP tasks that are being actively solved, propaganda techniques classification task is far behind, and both models and data are to be improved. However, nonzero scores for each class show that various baseline models are able to capture propagandist semantics and stronger models for this field can be built.

References

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*.

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

Violet Edwards. 1938. *Group leader’s guide to propaganda analysis: Revised edition of experimental study materials for use in junior and senior high schools, in college and university classes, and in adult study groups*. Institute for propaganda analysis, Incorporated.

Pankaj Gupta, Khushbu Saxena, Usama Yaseen, Thomas Runkler, and Hinrich Schütze. 2019. Neural architectures for fine-grained propaganda detection in news. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 92–97, Hong Kong, China, November. Association for Computational Linguistics.

Wenjun Hou and Ying Chen. 2019. CAUnLP at NLP4IF 2019 shared task: Context-dependent BERT for sentence-level propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 83–86, Hong Kong, China, November. Association for Computational Linguistics.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. *arXiv preprint arXiv:1910.02517*.

Hannah Rashkin, Eunsol Choi, Jin Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. pages 2931–2937, 01.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China, November. Association for Computational Linguistics.

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.

George-Alexandru Vlad, Mircea-Adrian Tanase, Cristian Onose, and Dumitru-Clementin Cercel. 2019. Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-capsule model. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 148–154, Hong Kong, China, November. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China, November. Association for Computational Linguistics.