Language Model, RoBERTa and T5 implementations for Detection of Propaganda Techniques in News Articles.

Natural Language Processing Project Proposal, Fall 2021

Antoine Basseto, Giacomo Camposampiero and **Andrea Pinto**
ETH Zurich - Swiss Federal Institute of Technology
{abasseto, gcamposampie, pintoa}@ethz.ch

1 Introduction

The proliferation of online misinformation has led to a significant amount of research into the automatic detection of fake news (Shu et al., 2017). However, most of the efforts have been concentrated on whole-document classification (Rashkin et al., 2017) or analysis of the general patterns of online propaganda (Garimella et al., 2018; Chatfield et al., 2015), while little has been done so far in terms of fine-grained text analysis. This approach could complement existing techniques and allow the user to extract more informed and nuanced judgment on the piece being read.

In this context, *Task 11 of SemEval-2020*[1] (Martino et al., 2020) aims to bridge this gap, facilitating the development of models capable of spotting text fragments where a defined set of propaganda techniques are used. This shared task provides a well-annotated dataset of 536 news articles, which enables the participant to develop detection models that automatically spot a defined range of 14 propaganda techniques in written texts (Martino et al., 2020).

The focus of the task is broken down into two well-defined sub-tasks, namely *(1) Span identification (SI)* to detect the text fragments representative of a propaganda technique in the news articles and *(2) Technique classification (TC)* to detect the propaganda technique used in a given text span.

2 Objectives and Goals

The primary goal of this Natural Language Processing project is two-fold:

1. To implement different models able to automatically detect the use of propaganda techniques in text snippets, accomplishing both the *Span identification* and *Technique classification* sub-tasks of the previously mentioned shared task. *SI* consists of a binary sequence tagging task, whereas *TC* consists of a multi-class classification problem (Martino et al., 2020).

2. To compare the implemented models and draw conclusions on their performance through an error analysis for each of them.

Furthermore, probings of possible future improvement tracks could be considered.

3 Implementation

Because the shared task is already a closed topic, many professional teams have made their way and produced a paper with satisfactory results. Martino et al. (2020) summarizes these results and outlines the winning teams' implementations. A literature review of these results was initially carried out to acquire information about possible implementation to be studied in our work.

As a result, the literature review outlined three interesting architectures that could be implemented to fulfill Goal (1).

1. A small self-trained language model to provide a baseline performance we can compare other models to. The model will be implemented using PyTorch (Paszke et al., 2019).

2. RoBERTa (Liu et al., 2019), a state-of-the-art pre-trained model based on the Transformer architecture; this model is particularly appealing as it has been used in most winning teams' ensembles (Chernyavskiy et al., 2020; Morio et al., 2020) with quite interesting results.

3. T5 (Raffel et al., 2020), another state-of-the-art Transformer based architecture that uses a text-to-text approach, whose performances could lead to good prediction performances. This particular model was also suggested to be of interest by the *aschern* team (Chernyavskiy

[1] The official task webpage: https://propaganda.qcri.org/semeval2020-task11/

et al., 2020), who achieved top scores in both sub-tasks.

To achieve Goal (2), the comparison between models will be carried out based on their performance in the shared task, using the F1-score on the test dataset as proposed by the task's organizers (Martino et al., 2020). A more thorough error analysis will be done for each model, pointing out any pattern in their weaknesses and possible ways of improvement for future work.

4 Milestones

The main deadlines outlined by the project regulation are reported in the following list.

- Project Proposal submission, *31/10/2021*
- Progress Report submission, *15/12/2021*
- Final Paper submission, *15/01/2021*
- Project Presentation, *18/01/2021*

Other than these important milestones, the project group will also carry out weekly meeting in order to ensure a good communication and a regular project progress during the course of the semester.

5 Expected results

Our team expects to be able to successfully implement all the three mentioned architecture and carry out an objective and complete error analysis on these models. Eventually, achieving good classification results on both *SI* and *TC* will be considered a great success, even if not the primary goal of this experience.

References

Akemi Takeoka Chatfield, Christopher G. Reddick, and Uuf Brajawidagda. 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, dg.o '15, page 239–249, New York, NY, USA. Association for Computing Machinery.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. aschern at semeval-2020 task 11: It takes three to tango: Roberta, crf, and transfer learning. *CoRR*, abs/2008.02837.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1).

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. *CoRR*, abs/1907.11692.

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. *CoRR*, abs/2009.02696.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online). International Committee for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.