HOMADOS at SemEval-2021 Task 6: Multi-Task Learning for Propaganda Detection

Konrad Kaczyński[1,2] and **Piotr Przybyła**[1]

[1]Institute of Computer Science, Polish Academy of Sciences
[2]Faculty of Economic Sciences, University of Warsaw
{konrad.kaczynski,piotr.przybyla}@ipipan.waw.pl

Abstract

Among the tasks motivated by the proliferation of misinformation, propaganda detection is particularly challenging due to the deficit of fine-grained manual annotations required to train machine learning models. Here we show how data from other related tasks, including credibility assessment, can be leveraged in multi-task learning (MTL) framework to accelerate the training process. To that end, we design a BERT-based model with multiple output layers, train it in several MTL scenarios and perform evaluation against the SemEval gold standard.

1 Introduction

Fine-grained propaganda detection is a new approach to tackling online misinformation, highlighting instances of propaganda techniques on the word level. These techniques are used in textual communication in order to encourage certain beliefs, but instead of straightforward presentation of arguments, they rely on psychological manipulation, logical fallacies or emotion elicitation.

There are general-purpose natural language processing (NLP) methods that could be used for automatic detection of such text fragments. The challenge here is that they require large amounts of training data, which are laborious to produce. However, propaganda techniques are often related to other misinformation challenges, for which large datasets do exist, e.g. credibility assessment or fake news detection.

In the present study we aim to investigate how this connection can be used in the multi-task learning (MTL) framework. We show how the performance of multi-label token-level propaganda detection within shared task 6 at SemEval-2021 can be improved by building neural architectures that are also trained to solve other tasks: single-label propaganda detection from SemEval-2020 and document-level credibility assessment based on a fake news corpus. We check different MTL scenarios (parallel and sequential) and show which aspects of the model benefit the most from this approach.

2 Problem Statement

We participate in SemEval-2021 Task 6 („Detection of Persuasion Techniques in Texts and Images"), subtask 2 (Dimitrov et al., 2021), where the goal is to identify all propaganda techniques within a provided fragment of text. Specifically, given a character sequence $\langle c_0, c_1, \ldots, c_N \rangle$, we aim to produce a set of annotations $\{(b_0, e_0, t_0), (b_1, e_1, t_1), \ldots, (b_k, e_k, t_k)\}$, where each triple consists of the character offsets of the span it covers ($0 \leq b_i < e_i \leq N$) and an indication which one from the set of 20 known techniques is detected there ($t_i \in T$). We can see it as a multi-label sequence classification task (Read et al., 2009), where each character (or token) can be assigned from 0 to 20 labels.

3 Related Work

Propaganda has been observed in text for a long time, but the problem of automatic detection of such techniques was posed just recently. Initially, a lack of word-level datasets confined the analysis to document-level classification, e.g. based on stylometric features (Rashkin et al., 2017; Barrón-Cedeño et al., 2019). Classification on the word level became possible with the dataset (Da San Martino et al., 2019b) released for the „Fine-Grained Propaganda Detection" shared task at the NLP4IF 2019 workshop (Da San Martino et al., 2019a). The corpus includes 550 news articles annotated with propaganda techniques on the word level. Among the submissions, the best performing models were based on word embeddings and pretrained lan-

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1027–1031
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

guage models, such as BERT (Devlin et al., 2018). To tackle the data sparsity problem, participants employed various over-sampling methods or trained their models on auxiliary data. Similar objectives were pursued at SemEval 2020 Task 11, consisting of two subtasks: binary sequence tagging task and multi-class classification task. The majority of tasks' participants based their solutions on the Transformers architecture (Vaswani et al., 2017) and word embeddings, combining them with other neural architectures (e.g. CNNs or LSTMs) or models such as CRF and logistic regression. Ensemble models were able to achieve satisfactory results when performing both tasks jointly.

4 Methods

4.1 Data Description

We make use of three datasets in English. In all the following approaches, the main focus is on the corpus released for SemEval-2021 Task 6 (Dimitrov et al., 2021) (S21). Additionally, we utilise the corpora from SemEval-2020 Task 11 (S20) (Da San Martino et al., 2020) and news credibility research (FN) (Przybyła, 2020).

S21 consists of text of 870 memes (607, 63 and 200 in the training, development and test subsets, respectively) annotated with 1550 spans a few words long (40 characters on average), each from one of 20 propaganda techniques. Most commonly occurring techniques are *Loaded language* (35%), *Name Calling/Labelling* (19%) and *Smears* (12%).

S20 corpus consists of 446 press articles (371 and 75 in the training and development subsets, respectively) annotated with 14 propaganda categories on a word level. Among the 7192 annotated spans, *Loaded language* (34%), *Name Calling/Labelling* (17%) and *Repetion* (12%) are most common categories. Given that very few spans overlap (8%), we represent the task as single-label classification by merging these spans according to their order in corresponding label files. Finally, we exclude sentences that do not contain any propaganda annotations.

To obtain the FN data, from the original corpus of 103,219 news articles classified as either credible or non-credible based on their source, we randomly select a sample of fifty thousand sentences with a binary credibility label.

4.2 Multi-Task Architecture

Figure 1 shows the architecture designed to fulfil the MTL objectives. A text document (usually one sentence) is first processed by BERT, resulting in 768-dimensional vectors: h_i for the i-th token and h_0 for the whole document, using the `[CLS]` token. These vectors are processed by classification modules D_x, each consisting of a linear dense layer and a softmax activation function. Three types of such operations are considered:

- $d_0 = D_d(h_0)$: document-level representation is used to produce 2-dimensional score vector (d_0), indicating class probabilities in binary single-label classification,
- $s_i = D_s(h_i)$: token-level representation is used to produce k-dimensional score vector (s_i), indicating class probabilities in multi-class single-label classification,
- $m_i = D_m(h_i)$: token-level representation is used to produce $l \times 2$-dimensional score vector (m_i), indicating class probabilities in multi-class multi-label classification.

The following subsections describe several scenarios of using these three output types to improve the accuracy of propaganda detection.

4.3 Single Task

In the primary method we use BERT-Base-Uncased with the token-level multi-label classification layer, trained using only S21 data (SINGLE S21). The output for the i-token, denoted by m_i, is a 20×2 matrix, in which the j-th row reflects the probability of the j-th propaganda technique being present in this token. If the token does not participate in any propaganda techniques, the first column of the matrix will be filled with ones and the second one with zeros. Since the S21 corpus is annotated at the character level, during preprocessing we map the initial annotation into tokens obtained via WordPiece tokenisation.

4.4 Sequential Multi-Task Learning

In case of sequential MTL, the main training described in previous section is preceded with training for one of two auxiliary tasks:

- single-label classification task on S20 corpus (MULTI-S S20-S21).
- document-level classification task with FN corpus (MULTI-S FN-S21).

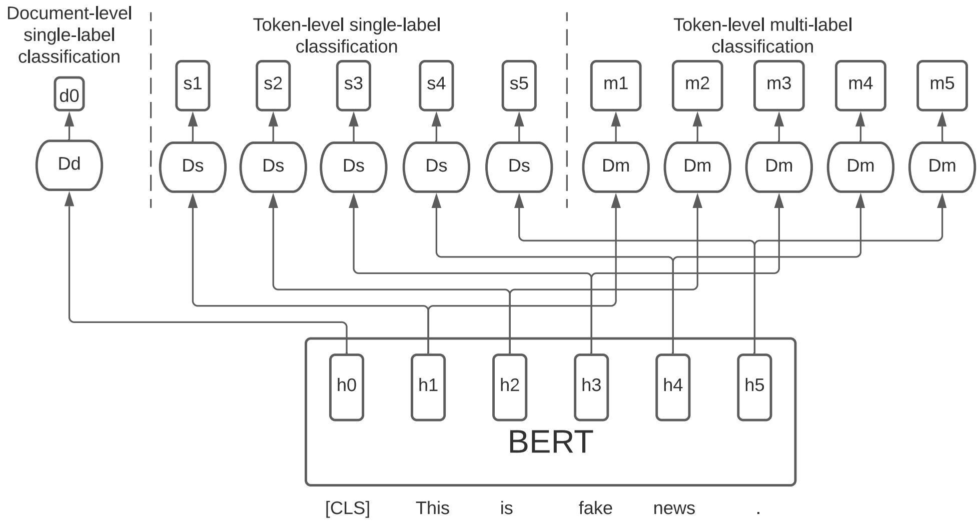
Document-level single-label classification
Token-level single-label classification
Token-level multi-label classification
d0
s1
s2
s3
s4
s5
m1
m2
m3
m4
m5
Dd
Ds
Dm
h0
h1
h2
h3
h4
h5
BERT
[CLS]
This
is
fake
news
.

Figure 1: Multi-task architecture of our solution

For MULTI-S S20-S21, we involve the token-level single-label classification layer to produce 16-dimensional s_i vector. This allows to classify each token in S20 corpus into one of 16 categories (14 propaganda + non-propaganda + padding). MULTI-S FN-S21 uses the document-level single-label classification output (d_0) layer for classifying sentences from FN corpus as coming from credible ($d_0 = (0, 1)$) or non-credible ($d_0 = (1, 0)$) articles.

For each model the learning procedure is the same: first, during an auxiliary task, only the additional classification layer is trained using cross-entropy loss and the auxiliary data. In the second phase, the training continues as a regular task on S21 data, as described in the previous section. Weights of all trainable variables are being updated in both phases.

4.5 Parallel Multi-Task Learning

In the parallel MTL objective, the auxiliary task and the target task are learnt jointly. Similarly to sequential MTL, we devise two models, each consisting of BERT with two separate classification layers on top:

- single-label and multi-label classification on S20 and S21 corpora (MULTI-P S20-S21),
- document-level and multi-label classification tasks on FN and S21 corpora (MULTI-P FN-S21).

Every batch of data consists of sentences coming from both datasets: four sentences from S20 or FN and four sentences from S21 are sent through their corresponding classification layers to produce outputs, and then count losses and update weights based on appropriate losses.

5 Evaluation

5.1 Experimental setup

We train our models according to multi-task scenarios, and use development subset of S21 to choose optimal number of training epochs of the final phase. The model trained up to this point is applied to test data to produce final predictions. In case of sequential MTL scenarios, this is preceded by training on additional corpora: on S20 for 10 epochs or on FN for 1 epoch. In case of parallel multi-task scenarios, the difference of training set sizes requires a special approach. For S20-S21, one epoch of training covers the whole S21 and 1/9 of S20. For FN-S21, we choose a balanced subsample of 18 thousand sentences and each training epoch covers the whole S21 and 1/30 of this subsample.

We use cross-entropy as the loss function, and compute it only for for non-padding tokens. For all experiments we use a maximum sequence length of 210 tokens, Adam optimizer (Kingma and Ba, 2015) with the learning rate of 3×10^{-5} and batch size of four sentences. We use the L1 regularisation (Ng, 2004) with α = 0.01. During fine-tuning of the model, weights of all trainable variables, including those in BERT, are being updated. During inference, we translate token-level labels back to character-level labels, including spaces and punctuation marks between adjacent tokens with identical labels. All experiments are conducted within the *TensorFlow* framework.

	Dev			Test		
Approach	F1	Precision	Recall	F1	Precision	Recall
SINGLE S21	0.5412	0.5798	0.5075	**0.4571**	0.4752	0.4403
MULTI-S S20-S21	0.5084	0.5181	0.4990	0.4444	0.4500	0.4390
MULTI-S FN-S21	0.4581	0.4836	0.4351	0.4185	0.4778	0.3723
MULTI-M S20-S21	**0.5455**	0.5747	0.5191	0.4074	0.4121	0.4028
MULTI-M FN-S21	0.5291	0.6429	0.4496	0.4381	0.5307	0.3730

Table 1: Propaganda detection performance on the development and test set for different evaluated approaches. The best F1 scores are highlighted. The run submitted to the shared task is underlined.

5.2 Evaluation measures

To evaluate our results we use character-level F1 measure prepared for the shared task (Dimitrov et al., 2021). It compares model's results with the golden annotations, accounting for the imbalance of categories and partial overlaps between fragments with the same label.

6 Results

Table 1 shows the performance of the considered approaches on the development and test set. The highest F1 score on the development set was obtained by the MULTI-P S20-S21 model. Hence, this model was used to generate the predictions on the test set submitted to the shared task (underlined). However, we can see that the single task approach is not far behind on the development set and actually provides the best performance on the test set. The differences between approaches are relatively modest and no single model outperforms others on each set and metric. This is mostly due to the small size of the propaganda datasets. Specifically, choosing the approach and number of training epochs based on the development set, which contains just 63 documents, may lead to overfitting.

In order to better understand how the introduction of MTL influences the models, we perform additional experiments. Firstly, in Table 2 we show F1 score for the recognition of each technique in single task and sequential MTL scenarios using both auxiliary datasets. One could expect the usage of S20, annotated with a similar set of propaganda techniques, to improve performance for overlapping labels, but the data do not confirm this. For example, the performance for the relatively large (12.7%) *Smears* (Smr) category improves noticeably, even though it was not present in S20. At the same time, we see F1 drop in case of some techniques present in both datasets, such as *Appeal to authority* (AtA) or *Slogans* (Slg). Clearly, the language constructions covered by these labels in case of press articles and memes are too different to offer clear advantage of MTL. The relationship with fake news detection is even weaker, resulting in many techniques not being recognised.

Technique	S21	M-S S20	M-S FN
AtA	0.6316	0.0000	0.7273
Atfp	0.0000	0.3966	0.0000
Bwf\D	0.6292	0.5824	0.0000
CO	0.0000	0.1667	0.0000
Dbt	0.0000	0.4578	0.1778
Ex-Min	0.4957	0.3221	0.5041
FW	0.3333	0.5397	0.0000
Gg	0.2222	0.0000	0.0000
LL	0.7038	0.6385	0.7135
NC-L	0.6136	0.6159	0.6830
Slg	0.3448	0.0000	0.3750
Smr	0.3839	0.5756	0.4743
Whtb	0.3830	0.0000	0.2222

Table 2: Per-technique F1 score on test set for different auxiliary datasets: S20 propaganda (S20) and fake news (FN) used in sequential multi-task scenario (techniques with no performance differences omitted for brevity).

Secondly, in Figure 2 we show how F1 on test set changes during training on S21 for the single task configuration and two scenarios based on S20 data: sequential and parallel. As expected, we see that pre-training allows our model to obtain good performance much faster, e.g. reaching F1=0.4 after 7 epochs instead of 14. But after longer training, the single-task approach catches up and beyond 20th epoch, when all version reach stable results, it outperforms the MTL variants.

7 Conclusion

In this work we explore how detection of propaganda techniques in text of memes can be facil-

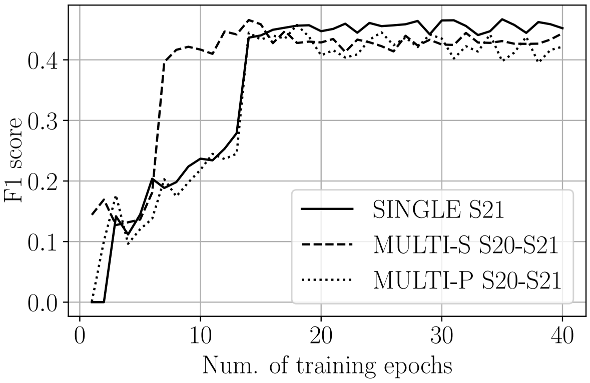

Figure 2: F1 scores during training for single task approach and multi-task learning using S20 data.

itated using external data in multi-task learning framework. The results show that the auxiliary tasks indeed influence the results, both in terms of accelerating the learning process and changing the set of recognised techniques. Nevertheless, these modifications do not offer clear advantages over the basic BERT-based solution.

We hypothesise this is because the link between main and auxiliary tasks is not strong enough to deliver benefits through multi-task learning. Additionally, propaganda is rarely a straightforward phenomenon and different techniques may require tailored approaches. We treat this effort as a preliminary study and aim to further investigate MTL's relevance in detecting propaganda by extending the auxiliary tasks portfolio with corpora reflecting other related issues, such as hate speech or hyperpartisan language.

Acknowledgements

This work was supported by the *Polish National Agency for Academic Exchange* through a *Polish Returns* grant number PPN/PPO/2018/1/00006.

References

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing and Management*, 56(5):1849–1864.

Giovanni Da San Martino, Alberto Barrón-Cedeño, 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, Hong Kong, China.

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 Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online).

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 the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2019)*, pages 4171–4186.

Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Task 6 at SemEval-2021: Detection of Persuasion Techniques in Texts and Images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, Bangkok, Thailand.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Andrew Y. Ng. 2004. Feature selection, L1 vs. L2 regularization, and rotational invairance. In *ICML '04: Proceedings of the 21st International Conference on Machine Learning*.

Piotr Przybyła. 2020. Capturing the Style of Fake News. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):490–497.

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 *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2931–2937.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5782 LNAI, pages 254–269.

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*, volume 2017-Decem, pages 5999–6009.