DUTH at SemEval-2020 Task 11: BERT with Entity Mapping for Propaganda Classification

Anastasios Bairaktaris **Symeon Symeonidis** **Avi Arampatzis**

Database and Information Retrieval research unit,
Department of Electrical and Computer Engineering,
Democritus University of Thrace, Xanthi 67100, Greece.
{anasbair1,ssymeoni,avi}@ee.duth.gr

Abstract

This report describes the methods employed by the Democritus University of Thrace (DUTH) team for participating in SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. Our team dealt with Subtask 2: Technique Classification. We used shallow Natural Language Processing (NLP) preprocessing techniques to reduce the noise in the dataset, feature selection methods, and common supervised machine learning algorithms. Our final model is based on using the BERT system with entity mapping. To improve our model's accuracy, we mapped certain words into five distinct categories by employing word-classes and entity recognition.

1 Introduction

According to the Institute for Propaganda Analysis[1], propaganda is an expression of an opinion or an action by individuals or groups deliberately designed to influence the opinions or the actions of other individuals or groups concerning predetermined ends. With the rapid change that the world wide web has made, it is evident that the means available for propaganda to be spread are more than ever before. The fact that, nowadays, news outlets can reach out to millions of people through their websites or social media demonstrates how easy it is to manipulate people with propaganda techniques or fake news. For example, political forecasts severely underperformed in predicting the results of the 2016 US presidential election and the United Kingdom European Union membership referendum (Brexit) as opposed to the consensus in social media, which is indicative of the new challenges that are upon us (Hall et al., 2018).

The SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles aims to produce models that can identify text fragments with various propaganda techniques. The first subtask is a binary sequence tagging task in which a model has to return the spans that contain at least one propaganda technique. The second subtask is a multi-class classification task in which given a text fragment and the article it occurs in, participants must classify the fragment into one of 14 different propaganda classes. More details on the Task can be found on the Task Description paper (Da San Martino et al., 2020).

The rest of this paper is structured as follows. Section 2 outlines some previous studies of propaganda identification. Section 3 describes our approach, while Sections 4 and 5 present experiments and results respectively. Conclusions are summarized in Section 6.

2 Background

Barrón-Cedeño et al. (2019) presented `Proppy`, a publicly available real-time propaganda detection system that is used for online news. The system used four modules that include article retrieval, event identification, deduplication, and propaganda index computation. To organize the news based on their propagandistic content, they showed that when identifying propaganda, approaches that use word n-grams are less effective than those that use character n-grams and other style features. Additionally, Da San Martino et al. (2019) introduced a new approach of analyzing propaganda that focuses on identifying

[1] https://propagandacritic.com/

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1732–1738
Barcelona, Spain (Online), December 12, 2020.

fragments that contain propaganda techniques as well as their type, as opposed to addressing propaganda detection at the document level.

Rashkin et al. (2017) described the need for examining lexical features when trying to understand the differences between more and less reliable digital news sources. They studied the usefulness of linguistic morphology in different types of fake news such as propaganda, satire, and hoaxes. They also created a corpus of categorized news articles with labels such as propaganda, trusted, hoax, or satire. In another major study, Rashkin et al. (2019) noted the importance of discovering relationships between different propaganda techniques. They hypothesized that finding common traits could prove helpful in classification tasks. In our approach, we investigated the effects of entity mapping in certain classes, and our conclusions are in line with Rashkin et al. (2017) concerning the existence of conceptual and linguistic relationships between propaganda techniques.

In recent years, there have been some significant landmarks in the NLP field. ELMo (Peters et al., 2018), ULMFiT (Howard and Ruder, 2018), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2018) are some large scale models that have massively improved the results in many NLP tasks. These systems provide models that have been pre-trained in massive corpora of unlabeled data and require fine-tuning in task-specific data. Although these systems offer excellent results, there is a need for further experimentation, as noted by Hua (2019) which highlights BERT's shortcomings in real-world scenarios.

3 Approach

This section describes our approach to mapping certain words into five distinct categories by employing word-classes and entity recognition. It also introduces the BERT model which was employed for our final submission.

3.1 Mapping the Dataset

The main idea of our method was to investigate the relationship between different entities and whether they have relevant usage. This is demonstrated with the examples in Table 1. The Flag Waving technique is an example of how words that bear no similarity in a bag-of-words representation, have the exact same semantic value for propaganda technique classification. In this example, 'Soviet Union' and 'Iran' have the same value (being both countries) for propaganda classification.

Propaganda Technique	Propaganda Extract
Flag Waving	*'This is not the **Soviet Union**, this is not **Iran** or Riyadh this is **America**.'*
Name_Calling,Labeling	*'**fascist** propaganda tropes.'*
Slogans	*'Make America Great Again'*

Table 1: Samples from different labels

The same applies for entities such as 'communists' and 'fascist' (political ideologies) and 'Christians' and 'Muslims' (religious groups). The hypothesis is that, for propaganda classification, when someone wants to attack another nation or a certain group through propaganda, it is less important which group or nation initiates or receives the attack. Thus, we made three lists that aim to reduce the noise in data that is produced from various countries, religious or political groups. We also made a list that contained different slogans to help with the Slogans category.

The lists we created are the following and can be found on github[2]:

- **List_Countries:** The names of 255 countries as well as some variations such as 'America' or 'UK'.
- **List_Religion:** 35 words that relate to religion such as 'Catholic' and 'Muslim'.
- **List_Politics:** 23 words that relate to politics such as 'Democrat' or 'Republicans'.
- **List_Slogans:** 41 slogans such as 'War on Terror' or 'Build the wall'.

[2] https://github.com/anasbair/SemEval2020-groups

We scanned the dataset for those instances and replaced them with the following tags: **NATION**, **RELIGION**, **POLITICS**, and **SLOGANS**. The final results showed that this approach improved significantly the basic BERT model.

3.2 Named Entity Recognition

Named Entity Recognition is the process of identifying proper names and classifying them into categories such as persons, organizations, locations, etc. This process is vital for many NLP applications (Petasis et al., 2001). Carrying on with our previous hypothesis, we also experimented with entity recognition. We noticed that in many instances of propaganda, there was a use of names of politicians that could be grouped to help the accuracy of our model. Although we experimented with many different entity groups/types such as Nationalities and Organisations, the best results came with the People entities.

To achieve this, we used SpaCy's[3] named entity recognizer which has been trained on the OntoNotes 5 corpus (Pradhan et al., 2007). After the recognition, we replaced the entity with the PERSON tag. This approach yielded our best results in the Flag Waving category.

3.3 BERT - Bidirectional Encoder Representations from Transformers

BERT is a language representation model that was introduced by Devlin et al. (2018). It stands for Bidirectional Encoder Representations from Transformers. BERT pre-trains deep bidirectional representations from text that has not been labeled. The fact that BERT is deeply bidirectional allows it to learn information during training from both sides of a token's context. The following two steps are involved in BERT.

The BERT model has been pre-trained in the BooksCorpus (800m words) (Zhu et al., 2015) and English Wikipedia (2,500m words). In the first step, we fine-tuned the BERT model on different versions of the dataset that was provided by the organizers. BERT requires input data to be in a specific format. To mark the beginning, the [CLS] special token is used and for the separation or end of the sentences the [SEP] is used. The input is represented as: $[\text{CIS}] + \text{text} + [\text{SEP}]$.

The next step was to tokenize the propaganda extracts into tokens that match BERT's vocabulary. For tokenization, we used BERT's `BertTokenizer`. `BertForSequenceClassification`[4] which is the model that we used for fine-tuning. This BERT transformer has a sequence classification/regression head on top (a linear layer on top of the pooled output). According to the recommendations of Devlin et al. (2018), for training we used a batch size of 32, a learning rate of 2e-5, and the number of epochs was set to 4.

4 Experimental Setup

In this section, we describe the experimental setup of this study, providing information for the dataset and the parameters of machine learning algorithms, respectively.

4.1 Dataset

The organizers provided three datasets Training, Development, and Test. The training dataset consisted of 357 articles in text format, retrieved with Python's newspaper3k[5]. For the second subtask, the organizers provided a text file with 6,129 propaganda text fragments, belonging to 13 categories, alongside their respective article id and the spans in which the technique was located in the article. The 13 categories/labels are shown in Table 4. The dataset is imbalanced since the Name_Calling,Labeling and Loaded_Language labels jointly constitute 50% of the dataset.

4.2 Pre-Processing

We tested various pre-processing techniques and by using the conclusions of Symeonidis et al. (2018) we applied the following: Remove Numbers, Remove Punctuation, Remove Symbols, Lowercase, and Replace all URL addresses normalizing them to 'URL'.

[3] https://spacy.io/

[4] https://huggingface.co/transformers/model_doc/bert.html#bertforsequenceclassification

[5] https://github.com/codelucas/newspaper/

We prefer not to remove stopwords due to the results of our previous work on SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums (Bairaktaris et al., 2019). In that work, we concluded that stopwords can prove important for certain tasks. For example, a common word such as ‘believe’ can strongly indicate opinion and as such is useful.

4.3 Machine Learning Model

Before using the BERT model for our final submission, we used standard machine learning methods for our experiments. We will briefly present these methods, which, just in two classes (Oversimplification and Flag Waving), performed better than the BERT model techniques on the development set.

For the training of our classifiers, we used Python’s Scikit-Learn library (Pedregosa et al., 2011). We split the given pre-labelled data into 2/3 training and 1/3 development set (2:1 ratio). After the split, the training set was shuffled, and tested a sequence of tuning parameters on the development set. When the test set was provided by Task organizers, we re-trained the classifiers into the total training set and tested on the organizers’ test set.

Vectorizer: We compared three common vectorizers such as CountVectorizer, HashingVectorizer, and TfidfVectorizer. Finally, our selection was the TfidfVectorizer since it yielded the best results.

Classifiers: We tested various classifiers and decided to use the following three: SGDClassifier, RidgeClassifier, and LinearSVC, as they yielded the best micro-averaged F_1 Results.

5 Results

This section summarizes our experimental results. Before our officially submitted run, we present some additional experiments.

5.1 Machine Learning Model results

In Table 2, we present the results of our machine learning Baseline Model. The Baseline Model is with the RidgeClassifier, as described in Section 4.3, since it yielded the best results on the training process. We show the F_1-score of the classifier when the Baseline Model was trained with the mapped datasets that we described in Sections 3.1 and 3.2. For the entity recognition we used a variety of entities such as persons, nationalities, organisations, countries, cities and locations. As we mentioned in Section 3.2 the PERSON entity achieved the best results.

The Baseline Model achieved some notable results on the development set for two labels. In the Oversimplification label, the baseline model yielded a micro-averaged F_1 of 29%, as opposed to the basic fine-tuned BERT which failed to recognize this class. Furthermore, for the Flag Waving label, the Baseline Model scored 1% more than our best BERT model. However, as we can see in Table 3, the BERT model performed better overall results and was selected for our final submission.

Technique	F_1 Score
Baseline Model (Overall)	46.37
NATION	47.13
RELIGION	47.13
POLITICS	47.22
SLOGANS	47.13
Combined Lists	47.03
PERSON	46.09
Various Entities	45.71

Table 2: Baseline Model performance on Development set

5.2 Bert Model Results

When fine-tuning the BERT model, we tried various approaches with the dataset. We tried using the raw dataset as well as a pre-processed one. Although pre-processing (with the techniques that we mentioned in Section 4.2) improved results over the raw dataset, when we applied the mapping and the named entity recognition techniques we observed that pre-processing did not help achieve better results. The results are presented in Table 3.

Technique	Development Set F_1
Baseline Model	46.37
BERT raw	51.14
BERT Pre-processed	56.44
BERT Various Entities	54.09
BERT Entity Person	56.91
BERT Entity Person Pre-processed	52.39
BERT Lists (all lists combined)	57.85
BERT Lists Pre-processed	55.03

Table 3: BERT effectiveness on different instances of the dataset

5.3 Final Submission Results

By examining the results of our BERT models, we concluded that the best results came with mapping the dataset with the NATION, RELIGION, and POLITICS labels. The second best approach was with the PERSON tag that outperformed our best model in the Bandwagon, Flag Waving, Labeling, and Cliches categories. Our official submission to the competition ranked our team to the 10th place from 32 teams. The results of our model are shown in Table 4.

Label	**Test Set F_1**
Loaded_Language	73.70
Name_Calling,Labeling	71.40
Repetition	20.10
Doubt	59.15
Exaggeration,Minimisation	28.23
Appeal_to_fear-prejudice	33.33
Flag-Waving	58.94
Causal_Oversimplification	26.23
Appeal_to_Authority	44.44
Slogans	34.78
Black-and-White_Fallacy	33.33
Whataboutism,Straw_Men	17.77
Thought-terminating_Cliches	27.02
Bandwagon,Reductio_ad_hitlerum	9.30
Overall micro-averaged F_1	57.20

Table 4: Final submission results

6 Conclusions

We presented a supervised learning model for classifying text fragments from news articles in thirteen propaganda categories. We used standard classification techniques as well as modern NLP models such as BERT. We examined the task from a sociological point of view and we tried to experiment with the fact that different entities of the same type can have the same value for propaganda classification. The results were promising and further experiments could improve them.

References

Tariq Alhindi, Jonas Pfeiffer, and Smaranda Muresan. 2019. Fine-tuned neural models for propaganda detection at the sentence and fragment levels. *CoRR*, abs/1910.09702.

Anastasios Bairaktaris, Symeon Symeonidis, and Avi Arampatzis. 2019. DUTH at semeval-2019 task 8: Part-of-speech features for question classification. In Jonathan May, Ekaterina Shutova, Aurélie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 1155–1159. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9847–9848. AAAI Press.

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. *CoRR*, abs/1810.04805.

Wendy Hall, Ramine Tinati, and Will Jennings. 2018. From brexit to trump: Social media's role in democracy. *IEEE Computer*, 51(1):18–27.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

Yiqing Hua. 2019. Understanding BERT performance in propaganda analysis. *CoRR*, abs/1911.04525.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Georgios Petasis, Frantz Vichot, Francis Wolinski, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D. Spyropoulos. 2001. Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France*, pages 418–425. Morgan Kaufmann Publishers.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Sameer S. Pradhan, Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2007. Ontonotes: a unified relational semantic representation. *Int. J. Semantic Computing*, 1(4):405–419.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

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 Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2931–2937. Association for Computational Linguistics.

Symeon Symeonidis, Dimitrios Effrosynidis, and Avi Arampatzis. 2018. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Syst. Appl.*, 110:298–310.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.