JUST at SemEval-2020 Task 11: Detecting Propaganda Techniques using BERT Pretrained Model

Ola Altiti
oaaltiti18@cit.just.edu.jo

Malak Abdullah
mabdullah@just.edu.jo

Rasha Obiedat
rmobiedat@just.edu.jo

Department of Computer Science
Jordan University of Science and Technology
Irbid, Jordan

Abstract

This paper presents the JUST team submission to semeval-2020 task 11, Detection of Propaganda Techniques in News Articles. Knowing that there are two subtasks in this competition, we have participated in the Technique Classification subtask (TC), which aims to identify the propaganda techniques used in specific propaganda fragments. We have used and implemented various models to detect propaganda. Our proposed model is based on BERT uncased pre-trained language model as it has achieved state-of-the-art performance on multiple NLP benchmarks. The performance result of our proposed model has scored 0.55307 F1-Score, which outperforms the baseline model provided by the organizers with 0.2519 F1-Score, and our model is 0.07 away from the best performing team. Compared to other participating systems, our submission is ranked 15th out of 31 participants.

1 Introduction

The high rate of using social media and the spread of digital news and online blogs have enabled the massive amount of data to reach a wide range of audience as well as allowed non-journalist to disseminate false news, misinformation, hoax and propaganda to mislead people and deceive them (Tandoc Jr et al., 2018; Rubin et al., 2015; Baisa et al., 2017). Moreover, the main way of accepting this news by society is the mass media and digital news(Gavrilenko et al., 2019). The reasons behind disseminating false information and news could be for financial and political purposes or to mislead the readers and influence their opinion negatively as well it has been argued to influence elections and threaten democracies (Shao et al., 2017; Abedalla et al., 2019). Propaganda is defined as"efforts by special interests to win over the public covertly by infiltrating messages into various channels of public expression ordinarily viewed as politically neutral (Sproule, 1994)". It aims to mislead audiences by influencing them toward a particular political or social agenda in news media (Volkova and Jang, 2018; Barrón-Cedeño et al., 2019). Therefore, several propaganda techniques and tools are designed to propagate certain ideologies. These techniques usually appeal to audience emotions and reach of their desires (Da San Martino et al., 2019).

Recently, a group of researchers (Da San Martino et al., 2019) proposed a fine-grained propaganda corpus and organized a shared task on fine-grained propaganda detection, which includes both sentence-level and fragment-level sub-tasks. Semeval task 11 is an extension to this task, which aims to predict propaganda techniques in news articles with two sub tasks:

- **Span Identification (SI):** Given a text article, identify the propagandist text spans
- **Technique Classification (TC):** Given a text span already flagged as a propagandist and its context, identify the specific propaganda technique it contains.

In this paper, we present our participation of the JUST team at semeval2020 task 11, and more precisely, we have participated in Technique Classification (TC) subtask. We have used a set of deep learning models

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1749–1755
Barcelona, Spain (Online), December 12, 2020.

with different word embeddings to detect propaganda techniques. However, the final submission was based on BERT's uncased pre-trained language model that achieved a significant performance. The rest of the paper is organized as follows.Section 2 describes the related work.Section 3 presents the dataset and per-processing.Section 4 presents the model and architecture. Section 5 presents the experiments and section 6 discusses the results.Finally, section 7 draws conclusions and sketches of future work.

2 Related Work

It is not surprising that people have been knowing and using propaganda for centuries (Shu et al., 2017; Petrov and Nakov,). For example, during World War One in 1914, propaganda was used on a global scale with the rise of the Nazi propaganda machine (Jewett, 1940) to mobilize hatred against the enemy. Detection of propaganda has gained massive interest in the research community in recent years. Therefore, the automatic detection of propaganda is studied as part of the propaganda analysis project in (Barrón-Cedeño et al., 2019). The researchers provided the first propaganda detection system, which is available publicly and is called poppy.This system is a real-world and real-time monitoring system to unmask propagandistic articles in online news. In (Gavrilenko et al., 2019), the researchers discussed the problem of identifying propaganda in online news content. Therefore, they performed several neural network architectures, such as Long Short-Term Memory (LSTM), hierarchical bidirectional LSTM (H-LSTM) and Convolutional Neural Network (CNN) in order to classify the text into propaganda and non-propaganda. Effective techniques were used for text pre-processing as well and different word representation models including word2vev, Global Vectors(GloVe), and TF-IDF. Moreover, the models performed to data provided from Twitter to the Internet Research Agency. The data was about the relevant activities of the IRA from September 1 to November 15, 2016. However, the results showed that CNN with word2vec representation outperformed other models with accuracy equals to 88.2 %. The researchers in (Da San Martino et al., 2019) released the shared task on Fine-Grained Propaganda Detection as part of the NLP4IF workshop at EMNLPIJCNLP 2019 that focused on detecting propaganda and specific propagandistic technique in news articles at sentence and fragment level, SLC and FLC tasks respectively. The wining system in SLC task (Mapes et al., 2019) was based on using an attention transformer using the BERT language model where the final layer of the model replaced with a linear softmax layer. However, to obtain multi-head results they have used ensemble attention neural networks with 12 attention heads and 12 transformer blocks.As well, the teams (Hua, 2019; Hou and Chen, 2019) fine-tuned BERT to tackle SLC task.Another team (Al-Omari et al., 2019) presented their proposed model to detect propaganda in the SLC task. Furthermore, they have experimented with various combinations of deep learning models including XGboost(Chen and Guestrin, 2016), BiLSTM and BERT cased and uncased with a set of features including effective words, lexical features as well as word embeddings based on Glove(Pennington et al., 2014). Moreover, their final model was based on an ensemble of XGboost, BiLSTM, and BERT-case and uncased where they achieved a 0.6112 F1 score. For the FLC task; a team (Yoosuf and Yang, 2019) achieved the best results by applying 20-way token level classification, each token in the input article classified into 20 token types. Moreover, they fined tune BERT uncased base model to classify the tokens by adding a linear head to the last layer of BERT.

3 Dataset and Preprocessing

The corpus which is provided by (Da San Martino et al., 2019) includes 350 articles for training and 75 articles for development. In training articles, there are 6129 propaganda spans and in the development set, there are 1063 propaganda spans. Figure 1 shows the distribution of classes in the training set, and as can be seen, the classes are imbalanced; loaded language is the most frequent class with 2123 samples whereas Bandwagon, Reductio_ad_hitlerum is the least frequent class with 72 samples. Text preprocessing has been performed for each span of training and development set that includes: removing punctuation, tokenization, cleaning text from special symbols, and cleaning contractions. All preprocessing steps were performed using the NLTK library.

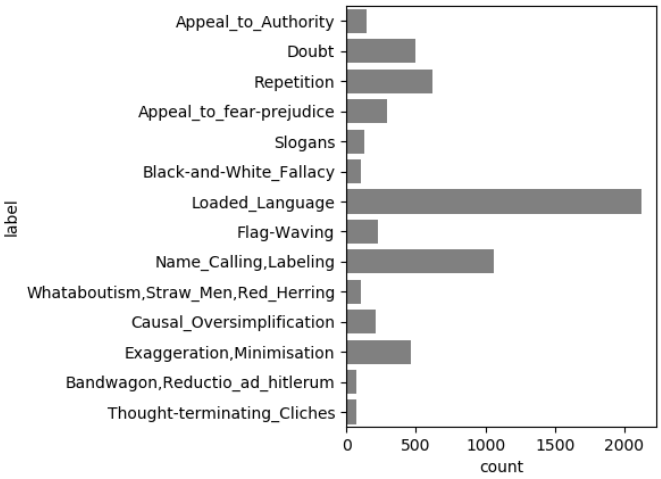

Figure 1: Classes Distribution for training set.

4 System Overview

We have used **Bidirectional Encoder Representations from Transformers (BERT)** (Devlin et al., 2018), which is one of the most powerful pre-trained language models that achieved state-of-the-art results in a wide variety of NLP tasks. There are two types of pre-trained models, BERT-Base and BERT Large, however, we have used the base model as it needs less computational resources. In each type, there are five pre-trained models, moreover, we have used Uncased and Cased. We have performed the pre-trained BERT models for multi-class classification, and retrieve bert-uncased, which contains 12 transformer layers and 110 million parameters as well bert-cased with 12 transformer layer, 768-hidden, 12-heads, 110M parameters.

BERT models have been downloaded from TensorFlow hub [1]. The architecture of the model is described in Figure 2. The cleaned span is represented using a special token [CLS] that is added in front of every span and another token [SEP] is added at the end. The input feeds the transformers' layers to generate the prediction for each class. Accordingly, the models are trained for 15 epochs using a learning rate of 2e-5, a maximum sequence length of 128, and a batch size of 128.

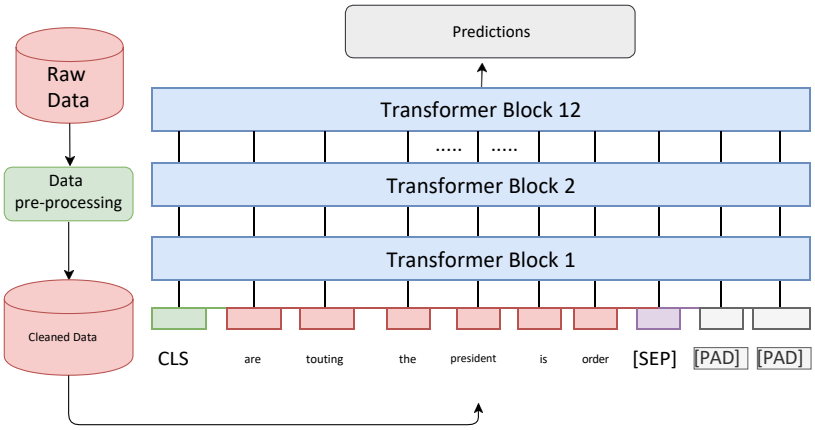

Figure 2: The architecture of our model using BERT.

[1] https://tfhub.dev/s?q=bert

5 Experiments

Additionally, we have experimented with several deep learning models based on learning representations (embeddings) that have been performed on the dataset to detect propaganda techniques. The input to the models is represented using a pre-trained sentence and word level embeddings to encode the input into the embedding vector. we used the word2vec embedding that is trained on Google News (Mikolov et al., 2013), where the sentence is encoded to embedding layers, which is a lookup table that consists of 300-dimensional pre-trained vector to represent each word. It is worth to mention that we have also experimented glove (Pennington et al., 2014) and fastText (Joulin et al., 2016) embeddings, but the results were not promising. For sentence embeddings, we have performed universal sentence encoder (USE) (Cer et al., 2018) that generates 512-dimensional representation for each sentence. The following deep learning models have been performed to the dataset:

- **Neural Network (NN) model**, the input is 512-dimensional representation for each span encoded using USE. This input passed to a fully connected neural network with three dense hidden layers of 128,128,75 neurons, respectively. A dropout of 0.4 and 0.2 have been added to avoid overfitting. The activation function for each layer is ReLU and for the output layer, one hidden dense layer used with the softmax activation function.

- **Convolutional Neural Network (CNN) model**, followed (Kim, 2014) architecture.The input to the CNN is a 300-dimensional representation for each input encoded using word2vec where the embedding is kept static. The second input is the 512-dimensional representation for the spans encoded with USE that is passed to a neural network with two dense hidden layers of 256 and 128 nodes, respectively, and dropout of 0.4. After that, the output from the first architecture concatenated with the output from the neural network passed to the dense output layer.

- **Bidirectional Long short-term memory (BiLSTM) model** (Hochreiter and Schmidhuber, 1997) with two different inputs that are fed into BiLSTM layers and Fully Connected Neural Network. The inputs encoded using word2vec embeddings and USE.

The models are implemented using Keras framework [2].USE loaded from TensorFlow hub[3].For training, we have used the Adam optimizer[4] with lr=0.001, and categorical Cross-Entropy as the loss function. The batch size is set to 32 and the number of epochs to 25.

6 Results and Discussion

We have directly evaluated all the models on the development set, and the best model is chosen to generate predictions of the test data. Table 1 shows the results of deep learning models, BERT-cased and BERT-uncased on the development set. The uncased model of the BERT language model gives the best prediction which applies that it works better than the cased model and outperforms other deep learning models. Hence, BERT able to understand the propaganda technique better than the other models. Also, we have tested the deep learning models with only word2vec embeddings and we have noticed that text classification using sentence embeddings (USE) outperforms word-level embeddings and provides a significant performance with minimal amounts of training data. This is expected since the word2vec embedding is context-independent and it does not encode the semantic relationships between words in the input sequence.

In the test stage, since we are only allowed to submit a single run on the test set, we choose the model with the highest F1 score on the development set (0.5766) to generate predictions, which is **BERT-uncased**. The evaluated results on the test set are listed in Table2, the model yielded a test F1 score of 0.55307.

Table3 shows the class-wise scores. Accordingly, the model performs well on propaganda techniques that appear frequently in the article such as "Load language"and "NameCalling, Labeling "that achieved

[2]https://keras.io/

[3]https://tfhub.dev/google/universal-sentence-encoder/1

[4]https://tfhub.dev/s?q=bert

Model	F1-score
Neural network with USE	0.531515
CNN	0.519285
BiLSTM	520226
BERT-uncased	**0.57667**
BERT-cased	0.556914

Table 1: Evaluation results on official development set

Rank	Team	F1-score
1	ApplicaAI	0.62067
2	aschern	0.62011
3	Hitachi	0.61732
4	Solomon	0.58939
...	...	..
15	**JUST**	**0.55307**
....	...	...
30	Baseline	0.25196

Table 2: Our results on the TC task alongside comparable results from the competition leaderboard.

Technique	Dev F1	Test F1
Appeal_to_Authority	0.27273	0.48889
Appeal_to_fear-prejudice	0.32967	0.37097
Bandwagon,Reductio_ad_hitlerum	0.44444	0.24490
BlackandWhite_Fallacy	0.30000	0.28889
Causal_Oversimplification	0.29268	0.27273
Doubt	0.53425	0.58258
Exaggeration,Minimisation	0.44737	0.29565
FlagWaving	0.72956	0.62564
Loaded_Language	0.73655	0.71958
Name_Calling,Labeling	0.69873	0.64727
Repetition	0.22886	0.21941
Slogans	0.50000	0.33333
Thoughtterminating_Cliches	0.12903	0.31818
Whataboutism,Straw_Men,Red_Herring	0.25000	0.28571

Table 3: Classwise F1 scores for final submission.

0.71958 and 0.64727 F1 score respectively on the test set, while BERT acts poorly on some propaganda types, such as "Exaggeration, Minimisation", "Bandwagon ", "Reductio_ad_hitlerum"and "Repetition". However, some propaganda techniques are challenging due to the way that the span is shaped or the number of words in the span, for example, "Flag-Waving"can be shaped in different ways, and "Repetition"depends on the occurrence of a word (or more) and the repetition of it in the article. Therefore, it is not enough to only look at the span to make the prediction, more information needs to be given to the model such as article context and word counts for each word in the span across the article thus including the whole article as the context is needed. Due to the limited time, we didn't experiment with the effect of adding these features. Finally, we have noticed that BERT has a strong performance in detecting some complex propaganda techniques while may perform poorly on other techniques and that because of the problem of imbalanced classes, which are impacting the performance of the minority class and leading the model to favor predicting the majority class.

7 Conclusion

In this paper, we described our solution in the propaganda technique classification subtask at SemEval-2020 task 11. We have investigated several models such as CNN, BiLSTM, and NN with word and sentence embeddings including word2vec and USE to detect propaganda techniques. However, our final solution was based on the BERT-uncased language model, which showed significant performance. The evaluations were performed using the dataset that was provided by semeval2020 task11 organizers.

Our proposed model is ranked in 15th place among 31 teams. Moreover, the F1-score that is achieved is 0.55307, which outperformed the baseline model (0.25196). The results confirmed that pre-trained language models(like BERT) are clearly a step forward for NLP and it has a strong performance in detecting propaganda techniques.

References

Ayat Abedalla, Aisha Al-Sadi, and Malak Abdullah. 2019. A closer look at fake news detection: A deep learning perspective. In *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, pages 24–28.

Hani Al-Omari, Malak Abdullah, Ola AlTiti, and Samira Shaikh. 2019. Justdeep at nlp4if 2019 task 1: Propaganda detection using ensemble deep learning models. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 113–118.

Vít Baisa, Ondrej Herman, and Ales Horák. 2017. Manipulative propaganda techniques. In *RASLAN*, pages 111–118.

Alberto Barrón-Cedeño, 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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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 5640–5650.

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*.

Olena Gavrilenko, Yurii Oliinyk, and Hanna Khanko. 2019. Analysis of propaganda elements detecting algorithms in text data. In *International Conference on Computer Science, Engineering and Education Applications*, pages 438–447. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Yiqing Hua. 2019. Understanding bert performance in propaganda analysis. *arXiv preprint arXiv:1911.04525*.

Arno Jewett. 1940. Detecting and analyzing propaganda. *English Journal*, pages 105–115.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Norman Mapes, Anna White, Radhika Medury, and Sumeet Dua. 2019. Divisive language and propaganda detection using multi-head attention transformers with deep learning bert-based language models for binary classification. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 103–106.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rostislav Petrov and Preslav Nakov. Fine-grained analysis of propaganda in news articles.

Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 83. American Society for Information Science.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, pages 96–104.

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

J Michael Sproule. 1994. *Channels of Propaganda.* ERIC.

Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

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.

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.