TUB at WANLP 2022 Shared Task: Using Semantic Similarity for Propaganda Detection in Arabic

Salar Mohtaj[1,2] and Sebastian Möller[1,2]

[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]German Research Centre for Artificial Intelligence (DFKI), Labor Berlin, Germany
{salar.mohtaj|sebastian.moeller} @ tu-berlin.de

Abstract

Propaganda and the spreading of fake news through social media have become serious problems in recent years. In this paper, we present our approach for the shared task on propaganda detection in Arabic in which the goal is to identify propaganda techniques in the Arabic social media text. We propose a semantic similarity detection model to compare text in the test set with the sentences in the train set to find the most similar instances. The label of the target text is obtained from the most similar texts in the train set. The proposed model obtained a micro-F1 score of 0.494 on the test data set.

1 Introduction

Social media has played a crucial role in recent year, having a great impact on different areas such as communication, entertainment, and politics. Beside their positive applications, social networks became an easily accessible medium to spread disinformation and propaganda in recent years. Based on Hamilton (2021), propaganda differs from mis/disinformation in that it need not be false, but instead, it relies on rhetorical devices which aim to manipulate the audience into a particular belief or behavior (Hamilton, 2021).

In this paper we present our proposed approach for sub-task 1 of the propaganda detection in Arabic social media text shared task (*WANLP* 2022). *WANLP* 2022 shared task includes two sub-tasks; identifying the propaganda techniques in tweets as a multi-label text classification task, and identifying the propaganda techniques used in tweet together with the span(s) of text in which each propaganda technique appears as a sequence tagging task. We only submitted results for the first subtask of the shared task. In this sub-task one or more propaganda techniques have been assigned to Arabic texts from social media (Alam et al., 2022). There are 21 propaganda techniques in the dataset that represent different approaches to manipulate the audience. More details about the different tasks can be found on the web-page of the shared task[1].

Although the first sub-task (identify the propaganda techniques) in a multi-label text classification problem, we used semantic textual similarity (STS) methods to identify related propaganda techniques to the instances in the test set. Based on the obtained results STS methods show competitive performance to the text classification models for the task.

The rest of the paper is organized as follow; Section 2 presents recent research on text based propaganda detection in social media. An overview of the data and the proposed approach are presented in Sections 3 and 4, respectively. We briefly review the obtained result and discuss it in Section 5. Finally, we conclude the paper and the system in Section 6.

2 Related Work

In this section we highlight some of the recent models for the task of propaganda detection using machine learning and Natural Language Processing (NLP) methods. The task of propaganda detection could be analyzed from two different perspectives; text analysis and network analysis perspectives (Martino et al., 2020). Here we focus on text analysis based models and review the recently developed models based on deep neural networks and pre-trained language models.

Vlad et al. (2019) proposed a model to detect propaganda in sentence level based on the BERT (Devlin et al., 2019) and an BiLSTM model. They formulated the task of propaganda detection as a binary classification task in which the model should distinguish propaganda and non-propaganda contents. The proposed model includes fine-tuning a pre-trained model on the task of emotion classification and feeding the output into a BiLSTM

[1]https://sites.google.com/view/propaganda-detection-in-arabic

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 501 - 505

architecture. The obtained results show that the model can significantly exceeds the baseline approach (Vlad et al., 2019).

In another study Vorakitphan et al. (2021) developed "protect" model for propaganda detection in text. As a propaganda detection pipeline, "protect" extract the text snippets from the input text, and then classify the technique of propaganda. The text snippets extractor module uses BERT pre-trained language model and feed the extracted text into the next step that is propaganda detection module. RoBERTa (Liu et al., 2019) pre-trained model is used for the classification task and propaganda detection.

As another research on propaganda detection as a text classification task, Barrón-Cedeño et al. (2019) proposed a binary text classifier based on different features includes readability level and writing style. They compared the performance of different supervised models such as logistic regression and SVMs for the task.

In addition to only text based models, some multimodal models have also been proposed in recent year to detect propaganda not only on text but also on images. As one of these efforts, a data set and a model are developed by Dimitrov et al. (2021) to propaganda identification in a multimodal setting. The compiled data set in this research contains 950 memes, each annotated with 22 propaganda techniques. It is collected from Facebook in includes different topics such as vaccines, COVID-19, and gender equality. They also proposed four different models, two unimodal and two multimodal models. For the unimodal setting, they used BERT and ResNet152 (He et al., 2016) for the text- and image-based models, respectively. The obtained results on the proposed data set show that the multimodal approach can outperform the unimodal training approaches.

In the proposed propaganda identification approach in this paper, we developed a model based on semantic similarity techniques unlike the highlighted researches in this section that formulated the task as a text classification problem.

3 Data

In this section we briefly describe the dataset for the first sub-task of *WANLP* 2022. The task organizers provided four data sets for sub-task 1 that includes train, development, dev_test and test sets. All the data sets are provided in *JSON* format that includes an id, the text of tweets and a list of labels for each instance.

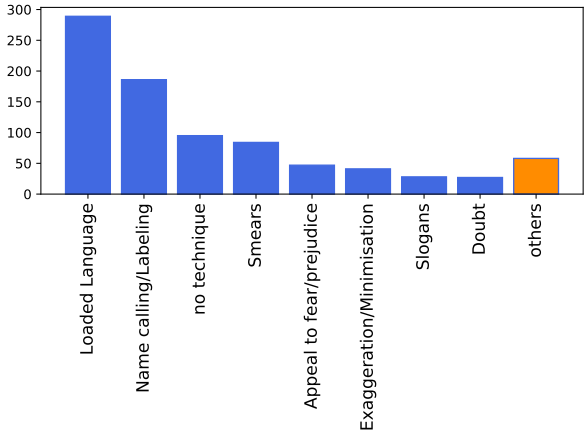

Figure 1: The distribution of top 8 most frequent propaganda techniques in the train set (504 data points). The "Others" represent the sum of the other techniques.

Table 1 highlights the main properties of the train and the test sets. As it is presented in the table, the instances in the test set tends to be shorter and also there are fewer number of words in the test set compared to the train data set.

There are a total number of 21 propaganda techniques in the data sets (Alam et al., 2022). Most of the techniques are presented and described in (Da San Martino et al., 2020). The distribution of ten most frequent techniques in the train set is depicted in Figure 1. As it is highlighted in the figure, "Loaded Language" is the most frequent propaganda techniques, followed by "Name calling/Labeling".

As the pre-processing step, we replaced twitter handles (i.e., the usernames) and URLs with constant texts ("username" and "weblink" respectively). These pre-processing steps have shown promising impact on the overal performance of related tasks like hate speech detection (Mohtaj et al., 2020). Although replacing URLs with the content of pages they refer to could be effective in related tasks like fake news detection (Mohtaj and Möller, 2022a), we decided to replace them with a constant text to prevent changing the length of input texts, drastically. The python regular expression package (re) has been used to replace the above mentioned texts in the data.

4 System

In this section we present the proposed model to detect the propaganda techniques in the social media

Attribute	Train set	Test set
Number of instance	504	323
Average length of instances (in words)	15.7	15.5
The length of longest instance (in words)	46	28
The length of shortest instance	5	6
Total number of words	7939	5273
Number of unique words	5069	3748

Table 1: The main properties of the train and the test sets.

text.

Although the first sub-task of the competition is a multi-label text classification task, we decided to use an STS based model to detect the most probable techniques for the instances in test set. From the provided definition for different propaganda techniques in (Da San Martino et al., 2020) and also from the provided data sets, lexicons play an important rule in a number of the techniques. For instance, "Loaded language" as the most frequent technique in the data set is defined as "using specific words and phrases with strong emotional implications to influence an audience" in (Da San Martino et al., 2020). Here, emotion lexicons are the main indicator of this propaganda technique.

Considering the role of lexicons and keywords on these techniques, using lists of related lexicons for each propaganda technique could show promising results on the task. However, due to lack of access to such resources in Arabic, we proposed an STS based approach to find the most similar instances in the train set to the unlabeled texts in the test set.

Word embedding models have shown promising performance on NLP tasks related to semantic textual similarity (e.g., plagiarism detection) (Asghari et al., 2019). However, it has been shown that the state-of-the-art contextual word embedding models (e.g., BERT) can outperform the traditional models in different NLP tasks like word similarity detection (Gupta and Jaggi, 2021) and hate speech and fake news detection (Mohtaj and Möller, 2022b). The overall proposed approach for *WANLP* 2022 to identify the most probable propaganda techniques for the instances in the test set is presented in Figure 2.

In the proposed model we used the Arabic BERT model (Safaya et al., 2020) to convert all the instances in the train and test sets into contextual vectors. We averaged word vectors to obtain the vector representation of sentences which resulted a vector with the length of 768 for each instance in the data sets. As the next step, we computed the cosine similarity between each instance in the test set (i.e., target sentence) with all of the instances in the train set. We took n most similar instance to the target sentence where the similarity is higher than a *threshold*. We named them as candidate sentences. Finally, top t frequent techniques in the candidate sentences have been chosen as the label of the target sentence. More details about the experiments are presented in Section 5.

5 Results and Discussion

In this section we briefly present our results based on the above mentioned model and discuss the main findings in the experiments.

As it is mentioned in the previous section, we tested three main hyper-parameters in our experiments; number of candidate sentences (n), minimum similarity threshold (*threshold*), and number of most frequent techniques to take from the candidate sentences (t). Table 2 summarizes the obtained results for different parameters on the validation set.

The Arabic BERT model has been used in all of the experiments to convert sentences to dense vectors. Although micro-F1 is the official evaluation metric that has been used by the shared task

Hyper-parameters			Macro F1	Micro F1
n	*threshold*	t		
20	0.4	3	**0.088**	0.475
10	0.4	3	0.072	0.459
5	0.4	3	0.065	**0.497**
5	0.5	3	0.058	0.435
5	0.5	1	0.036	0.298
10	0.5	1	0.037	0.310
5	0.6	3	0.009	0.050

Table 2: The obtained results by different hyper-parameters on the development set

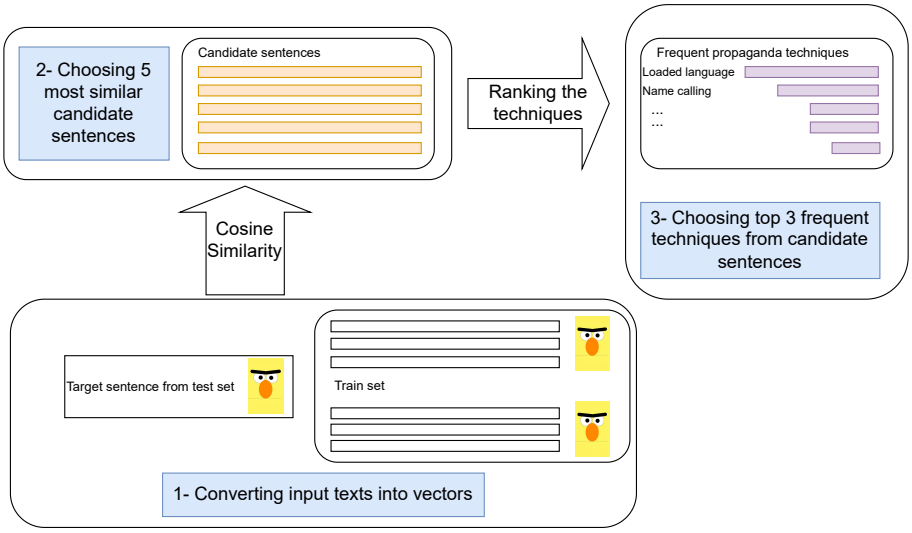

Figure 2: The overall process of choosing most probable propaganda techniques for a target sentence.

organizers to rank the models, the macro-F1 is also reported in the table. As it is highlighted in the table, the setting with 5 as the number of candidate sentences, 0.4 as the similarity threshold, and 3 as the number of top frequent propaganda techniques outperforms the other experiments. As a result, we submitted results on the test set based on this setting. The proposed model achieved the Micro F1 score of **0.494** and Macro F1 score of **0.076**.

One possibility to improve the overall performance of the proposed model would be using more than one hidden layer to convert the input text into vectors. In some studies on similar tasks like hate speech and fake news detection, it has been shown that using more than one layer for embedding could improve the performance of classification models (Mohtaj and Möller, 2022b).

6 Conclusion

In this paper we presented our model for the subtask 1 of the propaganda detection in Arabic social media text shared task (*WANLP* 2022). We proposed a model based on semantic textual similarity to compare instances in the test set with the labeled instances in the train set. The label of the test sentences obtained from the most similar sentences in the train set. Based on the results on the test set, the STS based model show a competitive performance compared to classification based models for the task.

For the future work, one can use different pre-trained language models to convert raw input texts into vectors. Moreover, the hyper-parameters can be tuned in order to improve the overall performance of the model.

Acknowledgements

We would like to thank the organizers of *WANLP* 2022 shared task for organizing the competition and taking time on the inquiries.

References

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Preslav Nakov, and Giovanni Da San Martino. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE. Association for Computational Linguistics.

Habibollah Asghari, Omid Fatemi, Salar Mohtaj, Heshaam Faili, and Paolo Rosso. 2019. On the use of word embedding for cross language plagiarism detection. *Intell. Data Anal.*, 23(3):661–680.

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

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). International Committee for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Prakhar Gupta and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5241–5253. Association for Computational Linguistics.

Kyle Hamilton. 2021. Towards an ontology for propaganda detection in news articles. In *The Semantic Web: ESWC 2021 Satellite Events*, pages 230–241, Cham. Springer International Publishing.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

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, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4826–4832. ijcai.org.

Salar Mohtaj and Sebastian Möller. 2022a. The impact of pre-processing on the performance of automated fake news detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 93–102. Springer.

Salar Mohtaj and Sebastian Möller. 2022b. On the importance of word embedding in automated harmful information detection. In *Text, Speech, and Dialogue - 25th International Conference, TSD 2022, Brno, Czech Republic, September 6-9, 2022, Proceedings*, volume 13502 of *Lecture Notes in Computer Science*, pages 251–262. Springer.

Salar Mohtaj, Vinicius Woloszyn, and Sebastian Möller. 2020. TUB at HASOC 2020: Character based LSTM for hate speech detection in indo-european languages. In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020*, volume 2826 of *CEUR Workshop Proceedings*, pages 298–303. CEUR-WS.org.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

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. Association for Computational Linguistics.

Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2021. PROTECT - A pipeline for propaganda detection and classification. In *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org.