IITD at the WANLP 2022 Shared Task: Multilingual Multi-Granularity Network for Propaganda Detection

Shubham Mittal[1] **Preslav Nakov**[2]
[1] Indian Institute of Technology Delhi
[2] Mohammed Bin Zayed University of Artificial Intelligence
shubhamiitd18@gmail.com, preslav.nakov@mbzuai.ac.ae

Abstract

We present our system for the two subtasks of the shared task on propaganda detection in Arabic, part of WANLP'2022. Subtask 1 is a multi-label classification problem to find the propaganda techniques used in a given tweet. Our system for this task uses XLM-R to predict probabilities for the target tweet to use each of the techniques. In addition to finding the techniques, Subtask 2 further asks to identify the textual span for each instance of each technique that is present in the tweet; the task can be modeled as a sequence tagging problem. We use a multi-granularity network with mBERT encoder for Subtask 2. Overall, our system ranks second for both subtasks (out of 14 and 3 participants, respectively). Our empirical analysis show that it does not help to use a much larger English corpus annotated with propaganda techniques, regardless of whether used in English or after translation to Arabic.[1]

1 Introduction

Propaganda is information deliberately designed to promote a particular point of view and to influence the opinions or the actions of individuals or groups. With the rise of social media platforms, the circulation of propaganda is even more pronounced since it may be built upon a true fact, but exaggerated and biased to promote a particular viewpoint. Various propaganda detection systems have been developed in recent years (Da San Martino et al., 2019; Barrón-Cedeño et al., 2019; Barrón-Cedeño et al., 2019; Dimitrov et al., 2021a,b), but they all have been restricted to English due to the unavailability of labelled datasets (containing fine-grained annotations of textual spans) in other languages. To bridge this gap, the WANLP'2022 shared task on propaganda detection in Arabic (Alam et al., 2022) released a dataset of Arabic tweets (we will call it ARATWEET) that uses 20 propaganda techniques, thus enabling research beyond English.

There are two subtasks defined in this shared task for detecting the propaganda techniques used in a tweet: (1) identify the techniques present in the given Arabic tweet, and (2) identify the span(s) of use of each technique along with the technique. Subtask 1 can be viewed as a multi-label classification problem, where the tweet may contain any subset of the 20 propaganda techniques, even all or none of them. Subtask 2 can be seen as a multi-label sequence tagging problem, where the system needs to predict the labels for each of the tokens. Subtask 2 is more challenging than Subtask 1 due to the increased level of detail it asks for.

Our Subtask 1 system uses a multilingual pre-trained language model, XLM-R (Conneau et al., 2020) to estimate a Multinoulli distribution over the 20 propaganda techniques for a given Arabic tweet. For Subtask 2, we use the multi-granularity network (MGN) from Da San Martino et al. (2019), but we replace the BERT encoder with mBERT (Devlin et al., 2019). We call our resulting system mMGN. Our systems, which use only ARATWEET data, rank second for both subtasks.

We investigated cross-lingual propaganda detection by using the Propaganda Techniques Corpus (PTC) (Da San Martino et al., 2019), which consists of annotated English news articles. We trained mMGN on PTC and continued its training on ARATWEET. Surprisingly, we found that continued training hurts the model by 10.2 F1 points absolute. To alleviate the possibility of ineffective transfer from English in mBERT embeddings, we further translated the PTC to Arabic using Google Translate and we projected the span-labels using `awesome-align` (Dou and Neubig, 2021). Upon doing continued training with a subset of the translated data, having only sentences containing propaganda, we found that it does not help, but also does not hurt the model. We believe that the domain difference between the two dataset is quite large, and thus there are no benefits in cross-lingual transfer.

[1] The code is released at github.com/sm354/mMGN

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 529 - 533

Propaganda Technique	train		dev		test	
	count	length	count	length	count	length
Appeal to authority	21	93.4 ± 43.9	8	94.8 ± 37.3	1	142.0 ± 0.0
Appeal to fear/prejudice	48	49.2 ± 29.0	11	54.9 ± 38.0	25	44.8 ± 27.9
Black-and-white Fallacy/Dictatorship	2	60.5 ± 12.5	3	56.3 ± 20.4	7	49.6 ± 19.8
Causal oversimplification	4	80.0 ± 43.2	2	57.0 ± 18.0	4	57.3 ± 24.2
Doubt	29	52.4 ± 34.6	3	61.0 ± 53.7	19	39.5 ± 21.3
Exaggeration/Minimisation	44	23.7 ± 28.4	26	14.3 ± 6.8	26	29.1 ± 16.9
Flag-waving	5	57.6 ± 30.7	4	65.0 ± 19.6	9	60.1 ± 23.2
Glittering generalities (virtue)	25	81.4 ± 48.9	9	66.1 ± 17.2	1	104.0 ± 0.0
Loaded language	446	9.70 ± 7.10	88	12.7 ± 13.2	326	7.20 ± 4.70
Misrepresentation of someone's position	0	N/A	0	N/A	1	37.0 ± 0.0
Name calling/Labeling	244	13.8 ± 6.4	77	15.6 ± 8.4	163	14.1 ± 6.6
Obfuscation, intentional vagueness, confusion	9	48.8 ± 28.1	4	34.0 ± 22.1	6	43.3 ± 23.6
Presenting irrelevant data (red herring)	1	61.0 ± 0.0	0	N/A	0	N/A
Reductio ad hitlerum	0	N/A	0	N/A	0	N/A
Repetition	9	12.8 ± 11.0	3	11.3 ± 4.1	3	35.3 ± 17.3
Slogans	44	17.0 ± 6.6	2	26.5 ± 13.5	6	24.5 ± 11.7
Smears	85	73.8 ± 34.9	27	88.8 ± 53.3	50	55.8 ± 22.0
Thought-terminating cliché	6	28.2 ± 17.5	2	21.0 ± 7.0	0	N/A
Whataboutism	3	47.7 ± 15.3	2	64.5 ± 20.5	0	N/A
Bandwagon	0	N/A	0	N/A	0	N/A
no technique	95	N/A	15	N/A	44	N/A

Table 1: Instance count of propaganda techniques and their span length in characters (mean ± std-dev) in the ARATWEET partitions. N/A is for either *no technique* or for those propaganda techniques having zero instances to compute mean/std-dev (such as *Misrepresentation of Someone's Position*, *Reductio ad hitlerum*, and *Bandwagon*).

2 Data

The dataset released in this shared task, which we call ARATWEET, comprises Arabic tweets, most of which (but not all) contain some propaganda techniques.

Table 1 shows statistics about the propaganda technique in the partitions of ARATWEET. Techniques such as *Misrepresentation of Someone's Position (Straw Man)*, *Presenting Irrelevant Data (Red Herring)*, *Reductio ad hitlerum*, and *Bandwagon* are rarely present in the dataset. *Loaded Language* is the most frequently present technique, whereas *Appeal to Authority* has the longest span. There are also tweets present that do not contain propaganda (e.g., 95 tweets in the training set).

Table 2 shows aggregated statistics about all propaganda techniques in the different partitions[2] of the dataset.

	train	dev	test
#examples	504	103	323
#spans	1025	271	647
tweet len (t)	15.8±6.1	18.6±9.9	15.4±5.0
tweet len (c)	112.6±39.2	123.4±58	117.4±30.6

Table 2: Statistics about the ARATWEET. Tweet len is the average length in # tokens (t) and # characters (c).

[2]The *dev* partition in this work refers to the combination of dev and dev_test released in the shared task.

3 System Description

Subtask 1 is a multi-label classification problem, where the model needs to find which of the 20 propaganda techniques (if any) are present in the input tweet. Our system (shown in Figure 1) fine-tunes a multilingual pretrained language model, XLM-R, (Conneau et al., 2020) for this subtask.

Given an Arabic tweet, we first tokenize it into word pieces $[T_1, T_2, \ldots, T_n]$ using the XLM-R tokenizer. We the pass these pieces through XLM-R to obtain contextualised embeddings, from which we take the *CLS* token embedding and we pass it through a single fully-connected linear layer to obtain a 20-dimensional embedding. After passing it through a sigmoid non-linearity, we convert this embedding, representing logits, to probabilities $[p_1, p_2, ..., p_{20}]$, one for each propaganda technique. Using a threshold of 0.5, our system assigns label i if $p_i \geq 0.5$. When $p_i < 0.5 \; \forall \; i$, the model predicts *no technique* for the target tweet.

Subtask 2 is a multi-label sequence tagging problem, where we want to label the tokens of a given tweet with the propaganda techniques. Since the (training) data contains tweets that do not contain propaganda (as discussed in section 2), we use the multi-granularity network (MGN) (Da San Martino et al., 2019) to develop our Subtask 2 system.

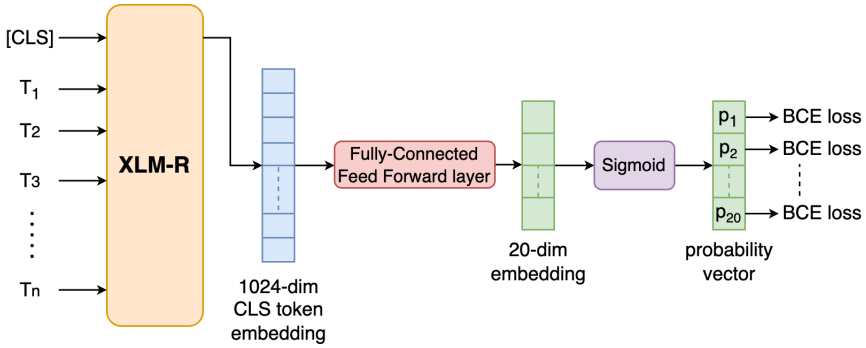

Figure 1: Our Subtask 1 system, which uses a pretrained XLM-R for multi-label classification. $T_1, T_2, T_3, \ldots, T_n$ are the tokens of the input tweet, and p_i is the probability of the tweet using i^{th} propaganda technique.

MGN uses BERT (Devlin et al., 2019) and models the task as a single-label sequence tagging problem, where either one of 20 techniques or none of them is assigned to each token. To improve the performance, it also adds a trainable *gate* to lower the probabilities for all tokens if the sentence does not contain propaganda.[3]

We replace BERT with mBERT in our MGN system, to obtain our multilingual multi-granularity network (mMGN) as our Subtask 2 system. mMGN can work for Arabic and for all other languages that are supported by mBERT.

4 Experiments

For evaluation, we use the official scorers that were released for the shared task. The official evaluation measure for Subtask 1 is micro-F1. However, the scorer also reports macro-F1. For Subtask 2, a modified micro-averaged F1 score is used, which gives credit to partial matches between the gold and the predicted spans.

We use the dev partition of ARATWEET to find the best model checkpoint and to report the scores on the finally released test set. Our models are trained on a single V100 (32GB) GPU.

Subtask 1 We empirically compare different pretrained language models (PLMs) as encoders for our Subtask 1 system and we report the scores in Table 3. With XLM-R encoder, our system achieves the best performance of 60.9 micro-F1. The hyperparameters of our Subtask 1 system include a maximum sequence length of 256, a batch size of 32, and 40 training epochs. We use two different learning rates: 1e-5 for PLM and 3e-4 for the remaining trainable parameters.

[3]We refer the readers to Da San Martino et al. (2019) for more detail.

	macro-F1	micro-F1
mBERT (Devlin et al., 2019)	8.1	54.3
AraBERT (Antoun et al., 2020)	**18.7**	59.4
XLM-R (Conneau et al., 2020)	18.3	**60.9**

Table 3: Performance(%) of our Subtask 1 system with different multilingual pre-trained LMs.

Subtask 2 We train the multilingual Multi-Granularity Network (mMGN) model on ARATWEET with a batch size of 16, a learning rate of 3e-5 for PLM and 3e-4 for other trainable parameters, and 30 epochs. This yields an F1 score of 35.5 on the test set, which is our best performance on this subtask.

Cross-lingual Propaganda Detection We ran several experiments using mMGN and the Propaganda Techniques Corpus (PTC), which is available in English (Da San Martino et al., 2019), to study cross-lingual transfer between English and Arabic in Subtask 2. In (1) ARATWEET, we train and test on ARATWEET, whereas in (2) PTC, we train on PTC data and we test in a zero-shot manner on ARATWEET. (3) TRANSPTC contains the translation of the PTC data from English to Arabic using Google Translate, followed by label projection using `awesome-align` (Dou and Neubig, 2021). Keeping only those translated sentences from TRANSPTC that contain propaganda gives (4) TRANSPTC+. (5) CTDTRANSPTC and (6) CTDTRANSPTC+ take the trained model from TRANSPTC and TRANSPTC+, respectively, and train it further on ARATWEET.

The performance across all settings is reported in Table 4. We can see that TRANSPTC is better than PTC by 0.6 F1 points, which suggests that the model learns better with the Arabic PTC.

	Precision	Recall	F1
AraTweet	35.5	25.7	**29.8**
PTC	53.1	1.4	2.8
TransPTC	30	1.8	3.4
TransPTC+	34.2	10.6	16.1
CtdTransPTC	21	18.4	19.6
CtdTransPTC+	30.6	28.0	29.2

Table 4: Performance(%) of mMGN (on dev_test) using different training methodologies.

The 1.8 recall of TransPTC is quite low, which could be due to the high proportion of propaganda-free sentences in PTC, which makes the model reluctant to propose propaganda techniques. When training only on propaganda-containing translated sentences from PTC, TransPTC+ improves over TransPTC on recall and also on precision, resulting in a gain of 12.7 F1 points absolute. Continued training on AraTweet, CtdTransPTC and CtdTransPTC+ yields sizable gains over the PTC-trained models TransPTC and TransPTC+. However, CtdTransPTC+ is worse than AraTweet by 0.6 F1 points absolute, indicating that cross-lingual transfer is not helping, but also not significantly hurting the performance.

We posit that the large domain difference between the PTC and the AraTweet datasets may be the reason for ineffective cross-lingual transfer. PTC contains news articles whereas AraTweet contains tweets, which causes linguistic differences in the text such as the presence of URLs, emojis, or slang in the tweets. Tweets are also often shorter due to text length limit in Twitter, which may also confuse the model between the two datasets.

5 Conclusion

We described our systems for the two subtasks of the WANLP 2022 shared task on propaganda detection in Arabic. For Subtask 1, we used XLM-R to estimate a Multinoulli distribution over the 20 propaganda techniques for multi-label classification. For Subtask 2, we used a multi-granularity network with mBERT, addressing the subtask as a sequence tagging problem. The official evaluation results put our systems as second on both subtasks, out of 14 and of 3 participants, respectively. We further described a number of experiments, which suggest various research challenges for future work, such as how to effectively use data from different domains, and how to learn language-agnostic embeddings for propaganda detection.

References

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France.

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*, pages 9847–9848, Honolulu, HI, USA.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 5636–5646, Hong Kong, China.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN, USA.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. 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*, pages 6603–6617.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b.

SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, pages 70–98, Online. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2112–2128.