SI2M & AIOX Labs at WANLP 2022 Shared Task: Propaganda Detection in Arabic, A Data Augmentation and Named Entity Recognition Approach

Kamel Gaanoun
SI2M Lab, INSEA
Rabat, Morocco
kgaanoun@insea.ac.ma

Imade Benelallam
SI2M Lab, INSEA
AIOX LABS
Rabat, Morocco
ibenelallam@aiox-labs.com

Abstract

This paper presents SI2M & AIOX Labs work among the propaganda detection in Arabic text shared task. The objective of this challenge is to identify the propaganda techniques used in specific propaganda fragments. We use a combination of data augmentation, Named Entity Recognition, rule-based repetition detection, and ARBERT prediction to develop our system. The model we provide scored 0.585 micro F1-Score and ranked 6th out of 14 teams.

1 Introduction

Even though the internet and social networks are tools for development and open doors to new opportunities, they also have a less attractive side. It is true that these tools are also used for bad purposes, such as the spread of propagandist messages when they are not identified as such by social media users. As part of cyber propaganda, or as part of the broader term "fake news" (Goswami, 2018), propaganda messages are used in social networks with the objective of convincing targeted populations in a biased way. Often, these messages aim to persuade their recipients to embrace ideas that are politically or ideologically motivated.

In light of the proliferation of such messages and the various upheavals the world is confronting today, researchers need to explore possible methods to detect cyber propaganda automatically. In contrast to English propaganda detection (Martino et al., 2020b), we note a flagrant lack of Arabic propaganda detection research, even if there are rare works dealing with this subject (Al-Ziyadi, 2019) or with close subjects like fake news (Nakov et al., 2022).

This work addresses this need, in order to build a system that can detect propaganda in tweets written in Arabic, as well as define the propaganda techniques employed. Indeed the dataset used in this paper contains 17 propaganda techniques, excluding "no technique", whose details are given by the organizers of the challenge (Alam et al., 2022) of which this work is part. Our system has the characteristic of combining a data augmentation method, Named Entity Recognition (NER), a rule-based approach, and the ARBERT model (Abdul-Mageed et al., 2020). The two main objectives are to answer the problem of the very limited amount of data available, and also to be able to detect as much as possible one of the most used propaganda techniques, namely "Name Calling/Labeling".

2 Related Work

Among the earliest definitions of propaganda is that of the Institute for Propaganda Analysis (Institute for Propaganda Analysis, 1938), which defined it in 1938 as "the expression of opinion or action by individuals or groups deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined end".

Apart from seeking the most comprehensive definition of the concept, several works have concentrated on categorizing propaganda techniques in order to better identify them. The first categorization was made by Clyde R. Miller (co-founder of the Institute for Propaganda Analysis) in 1937. Due to the proliferation of propaganda on social networks, these categorizations have become increasingly important over time due to the pressing need to detect propaganda automatically. The lack of annotated datasets dedicated to this problem, however, is one of the major obstacles. It was only in 2017 that the first datasets started to appear, namely the TSHP-17 (Rashkin et al., 2017), Qprop (Barrón-Cedeno et al., 2019) and PTC (Da San Martino et al., 2019b) in 2019.

In addition to detecting propaganda automatically, these datasets have also enabled us to detect the techniques in the texts in addition to specifying the relevant text fragments. Several works have emerged, mainly as system proposals within shared tasks. Like the Workshop on NLP4IF in

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 506 - 510
December 8, 2022 ©2022 Association for Computational Linguistics

2019 (Da San Martino et al., 2019a) and SemEval-2020 Task 11 (Martino et al., 2020a), both based on the TPC corpus. In the two shared tasks, two objectives were targeted simultaneously, namely the detection of the propaganda texts and the specification of the article part in question. The most effective solutions proposed can be summarized in the use of BIO encoding (Morio et al., 2020), self-supervision with the RoBERTa Model (Jurkiewicz et al., 2020) and BERT word-level classification (Yoosuf and Yang, 2019).

3 Data

We received two datasets from the challenge organizers, one named Train for training the system, and the other named Dev for validating and selecting the best configuration. The datasets contain a list of sequences and the propaganda techniques contained within these sequences. Also, at the end of the challenge, we receive a third dataset to evaluate the system. Using this last dataset, named Test, the final scores of each team are calculated. There is also a second task for which the same data is provided along with the start and end of the techniques within each sequence.

Table 1: Datasets content

Dataset	Number of sequences
Train	504
Dev	52
Test	323

Table 1 shows the number of sequences included in each dataset. Moreover, we note that the Train dataset contains 17 propaganda techniques, while the Dev dataset contains 16. We present the distribution of these techniques in Table 2. There is an over-representation of "Loaded Language" and "Name Calling/Labeling", followed by "Exaggeration/Minimisation" and "Smears", whereas the other techniques are very scarce, such as "Thought-terminating cliché", "Flag-waving", "Causal Oversimplification", "Whataboutism", "Black-and-white Fallacy/Dictatorship", and "Presenting Irrelevant Data (Red Herring)", which only occurs six times at most.

Table 2: Propaganda techniques distribution

Propaganda technique	Train	Dev
Loaded Language	446	46
Name calling/Labeling	244	44
Smears	85	12
Appeal to fear/prejudice	48	7
Exaggeration/Minimisation	44	10
Slogans	44	1
Doubt	29	1
Glittering generalities (Virtue)	25	7
Appeal to authority	21	7
Obfuscation, Intentional vagueness, Confusion	9	3
Repetition	9	2
Thought-terminating cliché	6	1
Flag-waving	5	2
Causal Oversimplification	4	1
Whataboutism	3	1
Black-and-white Fallacy/Dictatorship	2	1
Presenting Irrelevant Data (Red Herring)	1	0

Additionally, we present the most frequent combinations of techniques within the Train dataset sequences in Figure 1.

4 System

In the following sections, we describe our four-step system.

4.1 Data augmentation

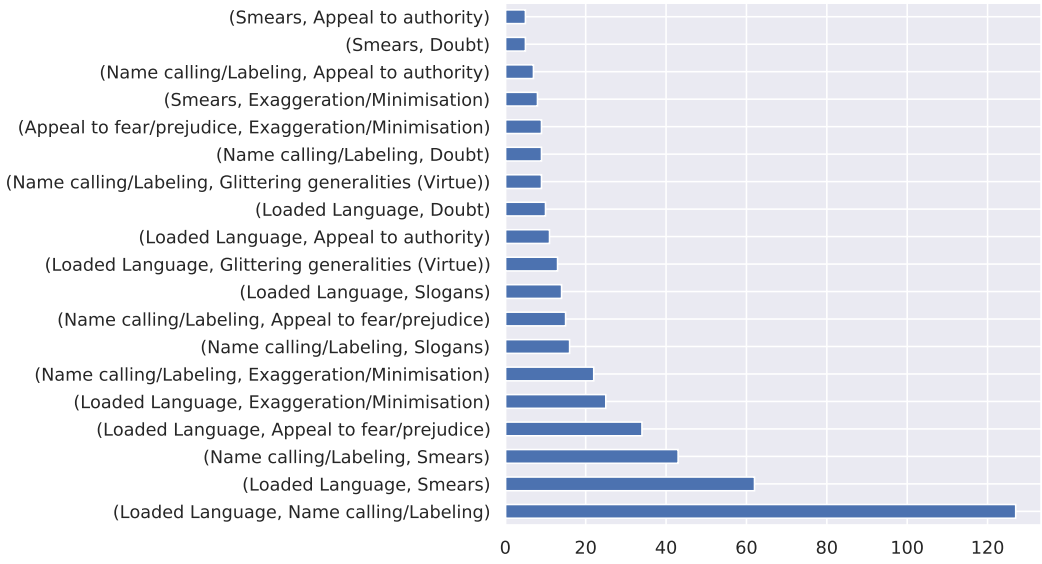
The first step is based on data augmentation. We use a strategy we call "MIX" adopted from (Gaanoun and Benelallam, 2020) work. The limited number of sequences available for training forces us to augment our data by generating synthetic sequences based on the mixture of subparts of the sequences we have. To do this, we take the following steps:

- Using Train and Dev sets including propaganda techniques tags (from second task data), we create a new dataset with one record per technique. The following is an example of retrieving two text chunks and their corresponding propaganda techniques:

  Sequence:

  [{'start': 1, 'end': 33, 'technique': 'Exaggeration/Minimisation', 'text': ده مش معتقل ده أحسن من اللوكاندة},
  {'start': 37, 'end': 86, 'technique': 'Smears', 'text': جدل وسخرية من زيارات تنظمها» «وزارة الداخلية للسجون}]

Figure 1: Most frequent propaganda techniques combinations in Train set



Produced records:

1. ده مش معتقل ده أحسن من اللوكاندة، Exaggeration/Minimisation
2. جدل وسخرية من زيارات تنظمها وزارة الداخلية للسجون Smears

- Synthetic sequences composed of two techniques are created by randomly mixing the produced sequences. The final system is based on a mixed dataset of 2000 examples. To evaluate on the Dev set, the Mixed dataset is concatenated with the Train dataset. After the better system has been validated, we concatenate the Mixed Dataset with both Train and Dev to evaluate it on the Test Dataset.

4.2 ARBERT prediction

ARBERT is fine-tuned based on our training data in a multi-label configuration, resulting in a list of detected techniques and their associated probabilities. Using these predictions, we retain techniques with a probability higher than a threshold defined using the Dev set. We evaluate the results of a list of thresholds and select the one that yields the highest micro-F1 score for the Dev set. We select 0.3 as our threshold for assessing the Test set.

When no technique has a prediction probability greater than the selected threshold, we label it "No technique".

Table 3 presents ARBERT training configuration and used infrastructure.

Table 3: AEBERT and infrastructure configuration

GPU	NVIDIA Tesla T4
Hyperparameters	Epochs: 20, batch size:8, learning rate:5e-5, Embedding maximum length: 512
Training average time	14 minutes

4.3 Named Entity Recognition

Name calling and labeling are frequently used in propagandistic messages to target an organization or a person. The goal of this type of propaganda is to engender a predefined feeling towards the object of the propaganda, whether it is a personality, an organization, a group, etc. We have therefore made the link with the detection of organizations or persons in the texts and the use of the NER method in order to better detect this technique. In order to accomplish this, we use a model pre-trained on the NER task(Sahyoun, 2022) based on the AraBERT model (Antoun et al., 2020). When this model detects the entity "ORG" in the text, we consider it to include the technique of "Name calling/labeling". The entity "PER" for the detection of the quotation of persons was also tested but did not give better results, it was thus abandoned for our final system.

4.4 Repetition detection

The repetition of words is one propaganda technique used to convince the recipient that the message is true. To improve the detection of this technique we use a rule-based method while removing the Arabic stopwords available through

the NLTK library. The repetition of one or two letter words is not considered in this step. Each time this method detects it and it is absent during ARBERT prediction, we add the "Repetition" technique.

Besides these 4 steps of the system, we also tried utilizing the PTC corpus (Da San Martino et al., 2020), which has the same purpose as the data used in this challenge, but is specific to English. Therefore, we proceeded to subtract the text chunks with their propaganda techniques. We then translated these chunks into Arabic using the Google Translate API[1]. Unfortunately, the use of this data did not improve the efficiency of the system, and was therefore not considered for further work.

5 Results

The results for the Dev and Test sets are presented in this section. To demonstrate the contribution of each of the steps considered in our system, we present the score obtained after applying each of these steps to the Dev set in Table 4. The final official results obtained on the Test set are presented in Table 5 .

Table 4: Dev set results for each step

Step	**micro F1**
Train set only	0.434
Mixed Data + Train set	0.455
Mixed Data + Train set + Repetition	0.459
Mixed Data + Train set + Repetition + NER (ORG)	0.56

Table 5: Official results on the Test set

micro F1	**macro F1**
0.585	0.137

We should point out that the official Test set result did not accounted for the label "No technique" in our predictions. This is because we used a capital N, whereas the organizers used a lowercase n for the final evaluation. The final result would have been 0.593 if this label had been considered.

[1]https://pypi.org/project/googletrans/

6 Discussion

The results show that the system's steps have a positive impact on the outcomes. Indeed, the score rises by 29% between the first step, which is solely based on the Train set, and the final step of the entire system. Furthermore, it appears that the use of NER has a significant effect on the final result, as the score knows the highest increase when using this method, recording a 22% increase. This finding is consistent with the fact that the name calling/labeling technique is the dataset's second most common technique. This result motivates future work to further investigate this idea by attempting to detect other majority techniques.

It is also worth noting that the data augmentation step contributed 5% to the improvement of the micro F1 score, whereas the detection of repetition contributed only 0.9%. The data augmentation step should be pushed in two directions: quantitatively by increasing the number of synthetic sequences generated, and structurally by prioritizing minority techniques or minority combinations in order to push the system to better predict these techniques.

7 Conclusion

This paper describes our contribution to the shared task of propaganda detection in WANLP 2022. We propose a system based on data augmentation, Named Entity Recognition (NER), repetition detection, and ARBERT prediction for subtask 1 dealing with multi-label classification techniques. Our analysis shows that NER and data augmentation have a significant impact on the final results, placing us sixth out of 14 competing teams.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Wafa Saeed Murshid Al-Ziyadi. 2019. *Propaganda-based Classification of Arabic Newspapers*. Ph.D. thesis, Hamad Bin Khalifa University (Qatar).

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

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

Giovanni Da San Martino, Alberto Barron-Cedeno, 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.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th Workshop on Semantic Evaluation*, SemEval '20, pages 1377–1414.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019b. 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 5636–5646.

Kamel Gaanoun and Imade Benelallam. 2020. Arabic dialect identification: An arabic-bert model with data augmentation and ensembling strategy. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 275–281.

Manash Pratim Goswami. 2018. Fake news and cyber propaganda: A study of manipulation and abuses on social media. *Mediascape in 21st Century: Emerging Perspectives*, pages 535–544.

Institute for Propaganda Analysis. 1938. How to detect propaganda. *Bulletin of the American Association of University Professors (1915-1955)*, 24(1):49–55.

Dawid Jurkiewicz, ukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. Applicaai at semeval-2020 task 11: On roberta-crf, span cls and whether self-training helps them. *arXiv preprint arXiv:2005.07934*.

G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at semeval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, et al. 2022. The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *European Conference on Information Retrieval*, pages 416–428. Springer.

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 *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Abdulwahab Sahyoun. 2022. arabert-ner. https://huggingface.co/abdusahmbzuai/arabert-ner. [Online; accessed 05-September-2022].

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.