iCompass at WANLP 2022 Shared Task: ARBERT and MARBERT for Multilabel Propaganda Classification of Arabic Tweets

Bilel Taboubi
bileltaboubi20@gmail.com

Bechir Brahem
bechir.brahem@outlook.com

Hatem Haddad
haddad.hatem@gmail.com

Abstract

Propaganda content has seen massive spread in the biggest social media networks. Major global events such as Covid-19, presidential elections, and wars have all been infested with various propaganda techniques. In participation in the WANLP 2022 Shared Task(Alam et al., 2022), this paper provides a detailed overview of our machine learning system for propaganda techniques classification and its achieved results. The task was carried out using pre-trained transformer based models: ARBERT and MARBERT. The models were fine-tuned for the downstream task in hand: multilabel classification of Arabic tweets. According to the results, MARBERT and ARBERT attained 0.562 and 0.567 micro F1-score on the development set of subtask 1. The submitted model was MARBERT which attained a 0.597 micro F1-score and got the fifth rank.

1 Introduction

Propaganda is one type of information that is shared deliberately for gaining political and religious influence. It is the systematic and deliberate way of shaping opinion and influencing the thoughts of a person for achieving the desired intention of a propagandist. In the age of "Post-truth" (Higgins, 2016), anti-scientific thinking and conspiracy theories the promotion of doctrines and ideologies that aim to manipulate and influence readers have rapidly spread through new communication mediums. In India, TV played a major role in the 2014 election, and some research has concluded that their results may have been swayed by propaganda techniques (Ward, 2014). Furthermore, social media platforms have known a widespread of propaganda, misinformation, and hate speech in their content. During the November 2012 Gaza conflict, Israel Defense Force and Hamas' Alqassam Brigades posted graphic images of death and suffering as well as explicit propaganda illustrations through their Twitter accounts(Seo, 2014). Social media platforms through their selective recommendation algorithms and their massive reach have fostered propaganda networks and "echo chambers" that amplify certain agendas and hide counter opinions and rebuttals. Propaganda actions may be now more effective than ever, representing a major global risk, possibly able to influence public opinion enough to alter election outcomes, decide wars, refuse Covid19 vaccines, and promote terrorism. For these reasons the need for modern automated and objective tools for uncovering propaganda is rising considerably.

2 Related Works

In the last few years research on detecting propaganda has seen a significant increase. The shared tasks found in workshops such as NLP4IF 2019(Yoosuf and Yang, 2019) and SemEval(Martino et al., 2020) (Semantic Evaluation) helped accelerate research on detecting propaganda and extracting the present propaganda techniques in a sentence or in a fragment of text. Also, apart from these workshops there exists work on binary classification of propaganda in the context of sentence-level and article-level(Oliinyk et al., 2020; Khanday et al., 2021).

On the other hand, Arabic propaganda detection research (Henia et al., 2021) is still lacking compared to its English counterpart (Taboubi et al., 2022). Our study presented in this paper attempts to classify propaganda techniques (multilabel classification) found in textual tweets using deep learning techniques and transformer architectures such as ARBERT and MARBERT.

3 Data

3.1 Data format and Characteristics

The data consists of a list of Arabic social media texts (tweets) and contained the list of propaganda techniques used in each tweet (table 1). The de-

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 511 - 514
December 8, 2022 ©2022 Association for Computational Linguistics

tails of the dataset are reported in (Alam et al., 2022) and we used the dataset of task 1. The pro-

id	text	labels
7365	تحذيرات من حرب جديدة في حال فشل الانتخابات القادمة	Loaded Language Appeal to fear/prejudice
7375	"بوليساريو يروج زيفا لصور من ليبيا تدعي أنها لطائرة مغربية تم إسقاطها"	Smears Name calling/Labeling

Table 1: Two samples from our data. "text" is a string containing the Arabic tweet textual data. "labels" are the propaganda techniques used in the tweet

paganda techniques used in the data are: Appeal to authority, Appeal to fear/prejudice, Black-and-white Fallacy/Dictatorship, Causal Oversimplification, Doubt, Exaggeration/Minimisation, Flag-waving, Glittering generalities (Virtue), Loaded Language, Misrepresentation of Someone's Position (Straw Man), Name calling/Labeling, Obfuscation, Intentional vagueness, Confusion, Presenting Irrelevant Data (Red Herring), Reductio ad hitlerum, Repetition, Slogans, Smears, Thought-terminating cliché, Whataboutism, Bandwagon, no technique. Our training dataset combined both files: "task1_dev.json" and "task1_train.json". This resulted in 556 data points. out of the 21 labels listed above, only 18 labels are present in the dataset. It is critical to note that the data is very unbalanced as some labels occur with orders of magnitude more than others. The propaganda technique "Loaded Language" was present in 346 tweets but "Presenting Irrelevant Data (Red Herring)" is present in only 2. This acute unbalance of the data is what pushed us to perform specific pre-processing methods on the data so that we give more chance to labels with lower frequency.

Figure 1 demonstrates the vast difference in the distribution of the labels' frequencies.

3.2 Data Preprocessing

Data preprocessing took the form of sequential steps listed here: remove emojis, normalization, remove links, remove special characters (i.e. ?,!,#), remove stop words.

4 System

To achieve the best results we have used language models (LMs) such as MARBERT and ARBERT (Abdul-Mageed et al., 2020). These models as their name suggest are based on the BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) language model which is a trained Transformer Encoder stack that uses bidirectional self-attention and was introduced by Google in 2018. While BERT focuses on the English language ARBERT and MARBERT were introduced to improve Arabic NLP tasks: ARBERT is pre-trained on standard Arabic language from sources such as Wikipedia and books. On the other hand, MARBERT focuses on dialectical Arabic. It is pre-trained on a large database of Arabic tweets On top of the BERT-based models, we have used global average pooling 1d and global max pooling 1d layers. Both of the pooling layers were concatenated and passed to a dropout layer and a final output layer.

We have tested both ARBERT and MARBERT with and without cross-validation. Cross-validation was done for 5 folds each with a percentage of 10% for the test. We also tested the models with and without the pooling layers. The training was done for 10 epochs using early stopping and we saved the best model on each epoch (according to the validation loss).

5 Results

We evaluated each model and each configuration at least 5 times and we calculated their mean and standard deviation. The results are plotted in (Fig 2)

F1 micro scores are presented in (table 2). From this table and its corresponding plot (Fig 2). From this table, we can see that the top 2 results are for ARBERT (without pooling and with cross-validation) mean: 0.567, std: 0.028 and MARBERT (with cross-validation and with pooling) mean: 0.562, std: 0.012. In the last two results, we have noted also their F1 macro scores: MARBERT mean:0.282, std: 0.023 and ARBERT mean: 0.243, std: 0.013

6 Conclusion

In this paper, we analyzed the performance of the pre-trained models ARBERT and MARBERT Despite the small-sized annotated data and huge unbalance presented in the provided data. To ob-

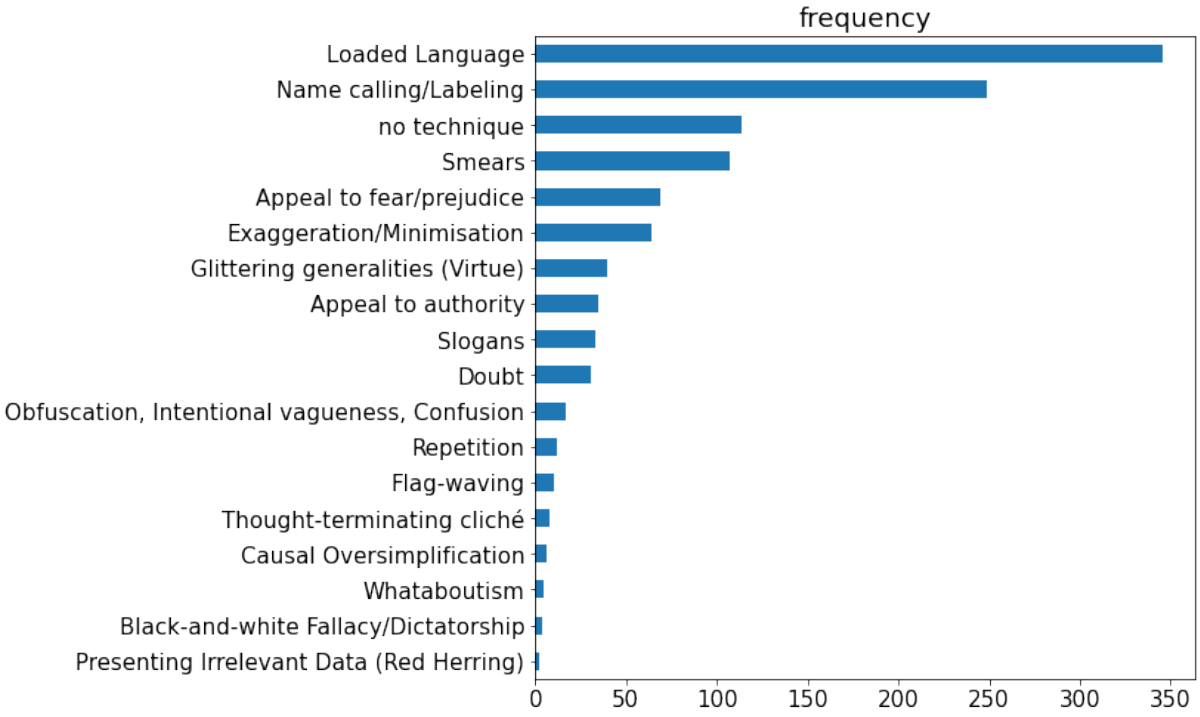

Figure 1: Horizontal bar plot that shows the distribution of the frequency of labels. It is clear from this plot that some labels are more present than others

with pooling		without pooling	
mean	std	mean	std
ARBERT with cross validation			
0.544	0.021	0.567	0.028
ARBERT without cross validation			
0.548	0.03	0.559	0.025
MARBERT with cross validation			
0.562	0.012	0.53	0.025
MARBERT without cross validation			
0.524	0.028	0.528	0.047

Table 2: Mean and standard deviation of the F1 micro score of the multiple runs for each model and training configuration

tain a good micro F1 measure for multilabel propaganda classification of Arabic tweets, different pre-processing techniques were applied to the data such as normalization, stopwords removal, etc. The submitted model MARBERT with pooling layers and trained with cross-validation splitting the data into 5 folds attained 0.597 for micro F1 and 0.191 macro F1 on the gold set reaching rank 5 on the leaderboard.

7 Limitations

Models attained unsatisfactory results for each of the micro and macro F1 measures and that is due to the low data distribution for many categories such as 'Whataboutism', and 'Black and white fallacy/Dictatorship'. Plus, the provided data was small in amount to train a model for multilabel classification with 18 categories. In the future, we will explore augmentation and resembling strategies to create a large balanced dataset for training and validating our proposed model and try to overcome our limitations.

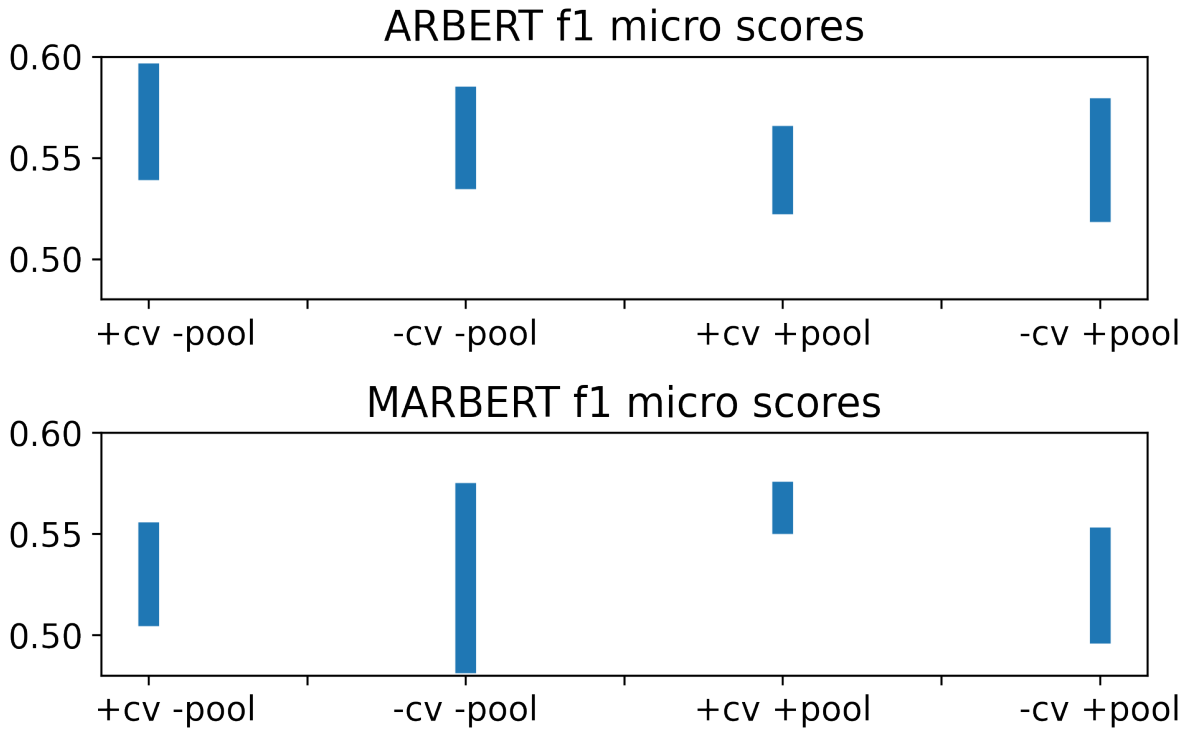

Figure 2: This figure plots the f1 micro scores and their errors for each training configuration. note that +cv (resp. -cv) means that the training was done with (resp. without) cross-validation. +pool (resp. -pool) means that the model used the two pooling layers (resp. did not use them)

References

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

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.

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*.

Wassim Henia, Oumayma Rjab, Hatem Haddad, and Chayma Fourati. 2021. iCompass at NLP4IF-2021–fighting the COVID-19 infodemic. In *Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda*, pages 115–118.

Kathleen Higgins. 2016. Post-truth: a guide for the perplexed. *Nature*, 540(7631):9–9.

Akib Mohi Ud Din Khanday, Qamar Rayees Khan, and Syed Tanzeel Rabani. 2021. Identifying propaganda from online social networks during covid-19 using machine learning techniques. *International Journal of Information Technology*, 13(1):115–122.

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

Vitaliia-Anna Oliinyk, Victoria Vysotska, Yevhen Burov, Khrystyna Mykich, and Vítor Basto Fernandes. 2020. Propaganda detection in text data based on nlp and machine learning. In *MoMLeT+ DS*, pages 132–144.

Hyunjin Seo. 2014. Visual propaganda in the age of social media: An empirical analysis of twitter images during the 2012 israeli–hamas conflict. *Visual Communication Quarterly*, 21(3):150–161.

B Taboubi, MAB Nessir, and H Haddad. 2022. icompass at checkthat! 2022: Combining deep language models for fake news detection. In *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*, pages 694–701.

Patrick Ward. 2014. Modi and the tv media: propaganda or profits? *ElEctIon*, page 53.

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.