AraBEM at WANLP 2022 Shared Task: Propaganda Detection in Arabic Tweets

Eshrag A. Refaee
Faculty of Computer Sciences
Jazan University
erefaie@jazanu.edu.sa

Basem H. Ahmed
Dept. of computer science
Alaqsa University
basem@alaqsa.edu.ps

Motaz K. Saad
Faculty of Info Technology
The Islamic University of Gaza
msaad@iugaza.edu.ps

Abstract

Propaganda is information or ideas that an organised group or government spreads to influence peopleś opinions, especially by not giving all the facts or secretly emphasising only one way of looking at the points. The ability to automatically detect propaganda-related language is a challenging task that researchers in the NLP community have recently started to address. This paper presents the participation of our team AraBEM in the propaganda detection shared task on Arabic tweets. Our system utilised a pre-trained BERT model to perform multi-class binary classification. It attained the best score at 0.602 micro-f1, ranking third on subtask-1, which identifies the propaganda techniques as a multilabel classification problem with a baseline of 0.079.

1 Introduction

With the increasing popularity of social media (SM) in our modern society, social media platforms like Twitter have become essential means for influencing others. People on social media tend to convey their views and perspectives more freely. The ability of SM users to freely express their views has allowed interested parties to profile users based on how they express their opinions on social media. As such, the past few years have witnessed a great interest in targeting a broader spectrum of audiences via Twitter and other SM platforms. Parties like political and advertisement campaigns are competing to reach out to the broadest audience base possible and hence influence the general publicś view. It can be seen that the more ability a particular part has to influence peopleś opinions, the more powerful it becomes (Ferrara, 2017).

The term propaganda is defined in the Cambridge dictionary as information or ideas that an organised group or government spreads to influence people's opinions by not giving all the facts or secretly emphasising only one way of looking at the facts.[1] The spread of propaganda exploits the anonymity of the Internet, the micro-profiling ability of SM platforms, and the power of automatically creating and managing coordinated networks of accounts to reach a large number of SM users with persuasive messages (Martino et al., 2020). Spreading propaganda to promote a specific agenda has become a business (Chatfield et al., 2015).

In this context, the concept of automatic propaganda detection has risen recently (Alam et al., 2022). Researchers have focused on utilising state-of-the-art NLP techniques to develop systems for automatic propaganda detection. The main challenges in this regard include difficulty identifying and extracting the linguistic signs of propaganda use. This is particularly difficult due to the cunning and indirect ways propaganda can be expressed. As such, detecting propaganda-related techniques can be challenging even for a human expert. Regarding Arabic, propaganda detection can be a more challenging task due to several additional factors, including the limited availability of linguistic resources (e.g., corpus) and the morphologically-rich nature of the Arabic language (Refaee, 2017).

To bridge this research gap, a shared task about auto-detection of propaganda in Arabic social media has been launched.[2] subtasks in this shared task and a comprehensive description of the shared task are discussed in (Alam et al., 2022). In this work, our team participated in the first subtask-1, ranking our system in the third position.

2 Related Work

The literature on propaganda detection as an NLP task reveals an increasing interest in exploring this

[1]The Cambridge Dictionary. Available at:https://dictionary.cambridge.org/dictionary/english/propaganda Accessed on 03/10/2022

[2]EMNLP-2022, SHARED TASK ON PROPAGANDA DETECTION IN ARABIC. Available at https://sites.google.com/view/propaganda-detection-in-arabic/home?authuser=0 Accessed on 02/09/2022

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 524 - 528

research area (Martino et al., 2020). Previous work on propaganda detection indicates several common challenges associated with this task. Specifically, the limited availability of the annotated dataset, the ability to convey propaganda with means other than text (e.g., images) (Hashemi and Hall, 2019) and the difficulty of spotting direct and indirect propaganda techniques are among the most prominent challenges of the task of propaganda detection. A total of eighteen propaganda techniques have been identified in previous work. However, experts stated that propaganda techniques are not fixed and keep evolving (Martino et al., 2020).

Previous work on propaganda detection has mainly focused on English (Chaudhari and Pawar, 2022), as a well-resourced language. In addition, researchers highlighted that most propaganda-related languages tend to appear in biased news and SM platforms, unleashing different directions like promoting political agendas and radicalisation (Albadi et al., 2019). (Chaudhari and Pawar, 2022) summarised the features utilised in existing systems for detecting propaganda techniques. This includes user-based, time-based, metadata-based and context-based features, n-grams, and pre-trained models (e.g., BERT).

In (Heidarysafa et al., 2020), the authors performed text mining on some of the propaganda content published by ISIS to recruit women from around the world. The authors applied a lexical-based emotion analysis method to detect emotions most likely to be evoked in readers of these materials.

Regarding propaganda detection in Arabic, literature shows that few previous attempts have been made to address this issue (Hashemi and Hall, 2019; Abozinadah et al., 2015; Albadi et al., 2019). This need has provoked the lunching of a shared task of propaganda detection in Arabic and releasing a newly built dataset annotated specifically for propaganda techniques (Alam et al., 2022).

3 Data

In this work, we utilise the dataset released for the shared task described in this overview paper (Alam et al., 2022). Table 1 shows the characteristics of the corpus. Our team performed cleaning up and pre-processing using the steps utilised in previous NLP tasks on Arabic (Refaee, 2017, 2021):

- Normalising exchangeable Arabic letters: mapping letters with various forms (i.e., *alef and Hamza* and *yaa*) to their representative characters (Antoun et al., 2020).
- Text segmentation: was performed to separate the tokens based on spaces and punctuation marks using the tokeniser provided by the PyArabic package (Zerrouki, 2010).[3]
- Removing diacritics, any special characters, punctuation, non-alphabetic characters and repeated characters, e.g., *loooool*.
- Normalising URLs, usernames, and hashtags.

Data	Training	Dev.	Testing
Size	504	52	323
# of Tokens	7792	747	4994
Avg. # of Tokens	15.46	14.36	15.46
# of Chars	51602	6436	34027
Avg. # of Chars	1102.38	123.76	105.34

Table 1: Size of the dataset split.

Class	Dist.
Misrepresentation of Someone's Position (Straw Man)	0
Reductio ad hitlerum	0
Presenting Irrelevant Data (Red Herring)	1
Black-and-white Fallacy/Dictatorship	2
Whataboutism	3
Causal Oversimplification	4
Flag-waving	5
Thought-terminating cliché	6
Repetition	7
Obfuscation, Intentional vagueness, Confusion	9
Appeal to authority	21
Glittering generalities (Virtue)	25
Doubt	27
Slogans	28
Exaggeration/Minimisation	41
Appeal to fear/prejudice	47
Smears	84
Name-calling/Labeling	186
Loaded Language	289

Table 2: Class Distribution in The Training Corpus

[3]PyArabic is a publicly available Python library explicitly designed for the Arabic language. Available at https://pypi.org/project/PyArabic/ accessed on 20/9/2022.

An initial observation reveals highly unbalanced classes in the obtained dataset, as shown in Table 2. Some classes have zero instances in the training set. Our team opted not to apply any technique to tackle class unbalancing. Instead, we decided to experiment with the original class distribution to explore how it would affect the overall system performance. It can also be seen that some propaganda techniques are more frequently occurring than others. For instance, the most commonly spotted propaganda techniques were *loaded language*, *name-calling*, and *labelling*. On the other hand, we noticed nearly a hundred tweets with no methods, which we decided to exclude from the dataset.

4 System

We explored approaches used in previous work on detecting propaganda or misleading language in social media. We noted that previous systems utilised different methods ranging from traditional machine learning (Habernal et al., 2017) to modern neural-based systems (Chetan et al., 2019).

Our team decided to use a pre-trained model, specifically BERT, which has performed well in previous work on propaganda detection in the news (Vlad et al., 2019; Badawy and Ferrara, 2018). The model has also been successfully used to detect auto-generated Arabic tweets, aiming to spot propaganda accounts (Harrag et al., 2021). We use BERT for multi-class binary label classification. The token size we use is 70 based on our calculation of the average tweet length. The output of running BERT on the shared-task dataset is several tensors, each associated with the possibility of the presence of propaganda techniques. To decide on the threshold, we ran several experiments and used the threshold 0.2, showing that any value above this threshold would be considered a presence of a propaganda technique. A possible future expansion of this work can include experimenting with a different threshold for each propaganda technique to test its impact on the overall system performance. We fine-tuned the pre-trained model using the training and development data.[4]

5 Results and discussion

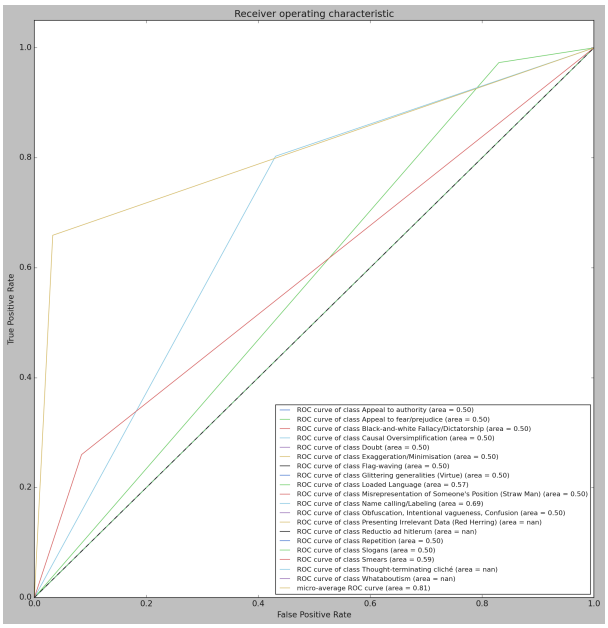
We used the script the shared task organisers provided to evaluate our system. The best results submitted by our system were reported at a micro F-1 score of 0.602, ranking third place in the leaderboard of the shared task. The details of all results can be found in (Alam et al., 2022). Overall, the scores attained by the participating systems reflect the difficulty of the task of auto-detection of propaganda language in Arabic tweets. We believe a possible explanation is a small size and highly unbalanced annotated dataset provided with the shared task. Identifying and annotating propaganda techniques can be challenging even for a human expert, (Panda and Levitan, 2021) and as such, expanding the scale of the dataset by using methods like data augmentation might help improve the performance. Another issue is that some propaganda techniques, are more frequently used than others, like *loaded language*. As mentioned in section 3 and table 2, some classes have zero or very few train instances. In contrast, others are either less regularly used or can be conveyed cunningly, making them hard to detect and identify. Overall, the results of the shared task indicate that more research is still required to identify misinformation in Arabic more accurately.

Class	P	R	F1
Appeal to authority	0.00	0.00	0.00
Appeal to fear / prejudice	0.00	0.00	0.00
Black-and-white Fallacy / Dictatorship	0.00	0.00	0.00
Causal Oversimplification	0.00	0.00	0.00
Doubt	0.00	0.00	0.00
Exaggeration / Minimisation	0.00	0.00	0.00
Flag-waving	0.00	0.00	0.00
Glittering generalities (Virtue)	0.00	0.00	0.00
Loaded Language	0.72	0.97	0.83
Misrepresentation of Someone's Position (Straw Man)	0.00	0.00	0.00
Name calling / Labeling	0.59	0.80	0.68
Obfuscation, Intentional vagueness, Confusion	0.00	0.00	0.00
Presenting Irrelevant Data (Red Herring)	0.00	0.00	0.00
Reductio ad hitlerum	0.00	0.00	0.00
Repetition	0.00	0.00	0.00
Slogans	0.00	0.00	0.00
Smears	0.36	0.26	0.30
Thought-terminating cliché	0.00	0.00	0.00
Whataboutism	0.00	0.00	0.00

Table 3: Precision (P), Recall (R), and F1-Score for each class label

[4]Access to the source code of our system is available on https://github.com/motazsaad/Arabic-Proagenda-Detection

Figure 1: Receiver Operating Characteristic ROC curve for each class label



Tables 3 and 4 show the Precision (P), Recall (R), and F1-Score for each class label and the average scores. Figure 1 shows the Receiver Operating Characteristic ROC curve for each class label. It is clear from the tables and the ROC curves that classes with more training instances have better results than the ones with few or zero training instances. The nature of this shared task is very challenging, and the scarcity of the dataset adds more challenges to it.

Metric	P	R	F1
Micro avg	0.6	0.6	0.6
Macro avg	0.1	0.1	0.1
Weighted avg	0.5	0.6	0.6

Table 4: Average Precision (P), Recall (R), and F1-Score for all classes

6 Conclusion

The occurrence of propaganda and misleading information to promote a specific agenda has coincided with the growing popularity of SM platforms like Twitter. Before that, news outlets would generally be the primary source of information for people; hence, the broadcast news would usually come trustworthy with no need to question or cross-check their legitimacy. Contrarily, the broad spectrum of audiences on SM platforms and the ability to readily access and spread propaganda to promote specific agendas has attracted interesting parties (e.g., political, advertisement, radicalization). As such, the need for NLP researchers to come together to address propaganda and fake news detection, especially in social media, has emerged.

In this context, a shared task has been launched to detect propaganda techniques in Arabic tweets automatically, and an annotated dataset has been released as part of the shared task. Our team, AraBEM, has participated in subtask-1, which is about classifying propaganda techniques, and ranked in the third position attaining a micro F-1 score of 0.602 compared to a baseline of 0.079. We used a pre-trained BERT model and decided on 0.2 as the threshold for tensors to determine if a propaganda technique was spotted. Overall, the results indicate a good performance among the participating team. However, more investigations are still required to enhance the system's ability to identify propaganda techniques accurately. Future directions for expanding this research include experimenting with different pre-trained models and threshold settings. In addition, more investigations on data balancing methods like data augmentation would shed light on distinct possibilities for performance improvement. More experiments should be done to assess the systems' ability to detect the more cunning and indirect ways of creating and spreading propaganda.

References

Ehab A Abozinadah, Alex V Mbaziira, and J Jones. 2015. Detection of abusive accounts with arabic tweets. *Int. J. Knowl. Eng.-IACSIT*, 1(2):113–119.

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.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2019. Investigating the effect of combining gru neural networks with handcrafted features for religious hatred detection on arabic twitter space. *Social Network Analysis and Mining*, 9(1):1–19.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for Arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Adam Badawy and Emilio Ferrara. 2018. The rise of jihadist propaganda on social networks. *Journal of Computational Social Science*, 1(2):453–470.

Akemi Takeoka Chatfield, Christopher G Reddick, and Uuf Brajawidagda. 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In *Proceedings of the 16th annual international conference on digital government research*, pages 239–249.

Deptii D Chaudhari and Ambika V Pawar. 2022. A systematic comparison of machine learning and nlp techniques to unveil propaganda in social media. *Journal of Information Technology Research (JITR)*, 15(1):1–14.

Aditya Chetan, Brihi Joshi, Hridoy Sankar Dutta, and Tanmoy Chakraborty. 2019. Corerank: Ranking to detect users involved in blackmarket-based collusive retweeting activities. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 330–338.

Emilio Ferrara. 2017. Contagion dynamics of extremist propaganda in social networks. *Information Sciences*, 418:1–12.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. *arXiv preprint arXiv:1707.06002*.

Fouzi Harrag, Maria Debbah, Kareem Darwish, and Ahmed Abdelali. 2021. Bert transformer model for detecting arabic gpt2 auto-generated tweets. *arXiv preprint arXiv:2101.09345*.

Mahdi Hashemi and Margeret Hall. 2019. Detecting and classifying online dark visual propaganda. *Image and Vision Computing*, 89:95–105.

Mojtaba Heidarysafa, Kamran Kowsari, Tolu Odukoya, Philip Potter, Laura E Barnes, and Donald E Brown. 2020. Women in ISIS propaganda: a natural language processing analysis of topics and emotions in a comparison with a mainstream religious group. In *Science and Information Conference*, pages 610–624. Springer.

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

Subhadarshi Panda and Sarah Ita Levitan. 2021. Detecting multilingual covid-19 misinformation on social media via contextualized embeddings. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–129.

Eshrag Refaee. 2017. Sentiment analysis for microblogging platforms in Arabic. In *International conference on social computing and social media*, pages 275–294. Springer, Cham.

Eshrag A Refaee. 2021. A data-oriented approach for detecting offensive language in Arabic tweets. In *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, pages 244–248. IEEE.

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.

Taha Zerrouki. 2010. PyArabic, an Arabic language library for python.