SIREN AI at WANLP 2022 Shared Task:
AraBERT Model for Propaganda Detection

Mohamad Sharara, Wissam Mahmoud, Ralph Tawil, Ralph Chobok, Wolf Assi and Antonio Tannoury

Abstract

Nowadays, the rapid dissemination of data on digital platforms has resulted in the emergence of information pollution and data contamination, specifically mis-information, mal-information, dis-information, fake news, and various types of propaganda. These topics are now posing a serious threat to the online digital realm, posing numerous challenges to social media platforms and governments around the world. In this article, we propose a propaganda detection model based on the transformer-based model AraBERT, with the objective of using this framework to detect propagandistic content in the Arabic social media text scene, with purpose of making online Arabic news and media consumption healthier and safer. Given the dataset, our results are relatively encouraging, indicating a huge potential for this line of approaches in Arabic online news text NLP.

1 Introduction

People are moving away from traditional media and toward digital content in today's landscape, and with trust in traditional media at an all-time low of 32% (according to a Gallup Inc. poll), it's no surprise that people are turning to alternative sources for news. Furthermore, social media has recently evolved into a major source of news content, giving rise to the "fake news" phenomenon (S. Shaden et al., 2021), in which a large amount of false information circulates, often with malicious intent (P. Nakov et al., 2021). The associated propaganda, which is almost always present in fake news, is an important but often overlooked feature of such destructive content (S. Yu et al., 2021). The primary goal of propaganda is to influence the opinions of target individuals through language manipulation (D. Dimitrov et al., 2021). There are over 313 million people worldwide who speak Arabic, with roughly 90% of them getting their news from the internet and online content. Furthermore, the prevalence of "fake news" in online content, as well as its amplification by social platforms, poses a number of serious challenges to society (D. Marc et al., 2020). Propaganda techniques, for example, Obfuscation, Black or White Fallacy, Loaded language, Name calling, Straw man, Red Herring, Whataboutism, and others, can pose grave threats to society, economy, democracy, health, journalism, the environment, and a variety of other areas.

While there have been recent studies that developed machine learning models to detect fake news in a variety of languages (N. Preslav et al., 2021), the lack of research into Arabic is, to say the least, concerning. Propaganda Detection in Arabic is a collaborative effort (F. Alam et al., 2022) to combat fake news by developing models for identifying propaganda techniques in Arabic social media text. So far, recent efforts to detect propaganda in news items around the world (G. Da San Martino et al., 2019) have addressed this as a fine-grained problem of finding it within fragments, and as a result, transformer-based embeddings work reasonably well in such detection approaches.

As a result, in this article, we attempt to achieve the goal of our contributions by following the flow:

- Data processing (given a small balanced dataset)
- Design a transformer model prototype oriented to Arabic propaganda detection
- Optimize the algorithm using the ADAM optimizer
- Examine and evaluate F1 score performance

2 Data

The dataset provided by the competition organizers (F. Alam et al., 2022) consisted of 504 Arabic tweets for training, each labeled by no more than five of several propaganda techniques. The no-technique label, which indicates that no propaganda technique was used in the tweet, was one of the labels. The majority of tweets were classified with one or two labels. In addition, a development set of 52 labeled tweets and a test set of 52 labeled tweets were provided. In terms of the total number of each label in the set and the number of labels per tweet, the latter two sets followed the same distribution as the training set. There was a total of 18 distinct labels, including the no-technique label. Labels included loaded language, Name calling, exaggeration, and so on.

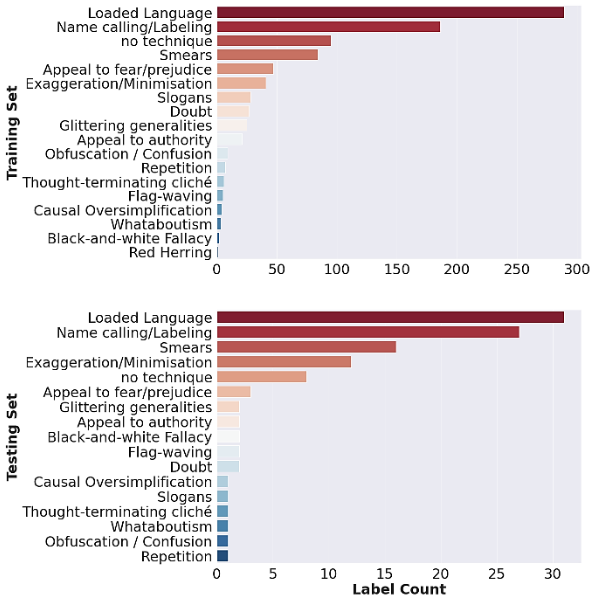

Figure 1: Label distribution across training and testing sets

The AraBERT model (W. Antoun et al., 2020) preprocessor was used to preprocess the training and testing sets. Preprocessing includes removing HTML markup, diacritics, tatweel, non-digit repetitions, and mapping Hindi numbers to Arabic, among so many other things. The highly skewed distribution of labels is a significant challenge in the competition. And, because the datasets are just so small, the model will be biased toward the most abundant label.

3 System

The model used for training is the AraBERT model's second version (W. Antoun et al., 2020). AraBERT is a powerful, cutting-edge transformer-based model for Arabic Language Understanding that has the same configuration as the base BERT model: twelve encoder blocks, twelve attention heads, seven hundred and sixty-eight hidden dimensions, five hundred and twelve maximum sequence lengths, and a total of approximately one hundred and ten parameters. The model included pre-trained embeddings that had been trained on approximately seventy million sentences from various sources.

The model was trained in the Google Collaboratory using a Tesla V100-SXM2-16GB GPU. The training set was divided into batches of sixteen each, with a gradient accumulation step of two, and an evaluation batch size of one hundred and twenty-eight. The optimization algorithm used is the ADAM optimizer, with an epsilon value of 10^{-8}, a learning rate of 0.00002, and iterates over twenty-five epochs. The overall training process took about five minutes, thanks to GPU parallelization. The model generates 18 probabilities, each of which corresponds to a propaganda technique, including no technique.

We used the following methodology to determine the output labels for each tweet based on those probabilities. We took the top five predictions from each tweet and discarded the rest. This is due to the fact that no tweet had more than 5 labels in the original data.

We had to optimize the threshold that will be used to filter out low probabilities from the top five. After some experimentation, we noticed that the most similar distribution was obtained with an optimal threshold of 0.35, assuming that the label distributions in the training, testing, and validation sets were the same.

We considered that if one of the remaining labels was ‘no-technique’, the corresponding tweet’s labels would be all labels with a probability greater than that of the no-technique label. Otherwise, the tweet will be labeled as no-technique.

We merged the three labeled datasets before predicting the labels of the unlabeled datasets and submitting the samples after training, validating, and testing on the datasets and reaching the optimal configuration.

4 Results

Because the competition organizers set the Micro-F1 score as the primary metric to evaluate the performance of the models, we used it to examine our model performance. This is primarily due to the fact that our task is simply to maximize the number of correct predictions made by the classifier, and no class is more important than the other. In the table below, we show the experimental results of applying AraBERT to the multi-label classification problem for Arabic propaganda detection:

Metric	Training loss	Validation loss	Macro-F1 score	Micro-F1 score
Score	0.19	0.3	0.108	0.4108

Table 1: Performance of our developed model on the test dataset

It is worth noting that we were able to achieve a Micro-F1 of 0.61 while using data augmentation and attempting to optimize the classification layer weights. Due to other deadlines, this result was not submitted. A Micro-F1 score of 0.578 on the evaluation dataset was a very promising result that will be improved using the methods described in the following section.

5 Discussion

Given that the Micro-F1 score on the training set is around 0.88, it is clear that the training data is overfit. The used model is cutting-edge and does a good job of capturing the data's complexity. However, the true issue is that the dataset is not representative enough for the model to generalize outside of it.

To improve the model's performance outside of the training set, the first step would be to add a regularization to the model, such as L2 regularization, to reduce overfitting. However, given the small dataset size, this may not be so promising. The second optimization approach would be to augment the existing data by performing some transformation on the documents, such as random insertion, random deletion, and word swapping. The latter method will increase data diversity at a lower cost than collecting brand new labeled data. A third approach that could be used is to use class weights to compensate for class imbalances by penalizing misclassification by infrequent classes (flag-waving, for example) more than that of more abundant classes. Given enough time, the most time-consuming but effective thing that could be done is to collect or manually label more propaganda-related tweets so that we have more representative data for real-world tasks.

6 Conclusion

In this paper, we presented a transformer-based model that serves as a contribution framework to identify propaganda types in Arabic text social media content (tweets basically), by highlighting the propaganda strategies utilized (such as Obfuscation, Black or White Fallacy, Loaded language, Name calling, Straw man, Red Herring, Whataboutism, and others).

With a Micro-F1 score of 0.578 and given the relatively small dataset, the model appears promising, and we are confident that performance improvements can be expected with a more balanced and richer dataset.

We intend to improve the model in the future by focusing more on the labeled dataset and expanding it by either applying careful, well-structured augmentation to some data or by developing a platform to assist annotators in labeling data. This ensures that the model is constantly updated and improved. Furthermore, we intend to conduct extensive research on various aspects of propaganda in order to develop a general propaganda detection system, thereby broadening the scope of our work in relation to the existing platform, with the goal of making the online Arabic journey healthier and safer.

7 Acknowledgments

This work was supported by Siren Analytics. We thank our colleagues who provided insight and expertise that greatly assisted the research. We want to thank our senior AI engineers, Wissam Antoun and Fady Baly who provided insights and expertise that greatly assisted the research and implementation phase. We also want to thank Dr. Elie Badine, our AI Lead, for his advice and observations, which helped to improve this paper.

References

P. Nakov and G. Da San Martino, *“Fake News, Disinformation, Propaganda, and Media Bias,” in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, p. 4862–4865.*

S. Yu, G. Da San Martino, M. Mohtarami, J. Glass, and P. Nakov, *“Interpretable Propaganda Detection in News Articles,” in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, p. 1597–1605.*

Alam Firoj, Mubarak Hamdy, Zaghouani Wajdi, Nakov Preslav and Da San Martino, Giovanni *“Overview of the WANLP 2022 Shared Task on Propaganda Detection in Arabic.” Proceedings of the Seventh Arabic Natural Language Processing Workshop, Association for Computational Linguistics, 2022.*

G. Da San Martino, S. Yu, A. Barrón-Cedeno, R. Petrov, and P. Nakov, *“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), 2019, p. 5636–5646.*

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

Shaar Shaden, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. *"The role of context in detecting previously fact-checked claims." arXiv preprint arXiv:2104.07423 (2021).*

Nakov Preslav, David Corney, Maram Hasanain, Firoj Alam, and Tamer Elsayed. *"Automated Fact-Checking for Assisting Human Fact-Checkers." in IJCAI, 2021.*

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov and Giovanni Da San Martino, *Detecting Propaganda Techniques in Memes, ACL, 2021.*

Djandji Marc, Antoun Wissam, Fady Baly, and Hazem Hajj. *" Multi-Task Learning using AraBert for Offensive Language Detection." in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, p.97-101.*