SussexAI at ArAIEval Shared Task: Mitigating Class Imbalance in Arabic Propaganda Detection

Mary Fouad
University of Sussex
Brighton, BN1 9RH, UK
University of Minia
Minia, Egypt
mm2157@sussex.ac.uk

Julie Weeds
University of Sussex
Brighton, BN1 9RH, UK
juliewe@sussex.ac.uk

Abstract

In this paper, we are exploring mitigating class imbalance in Arabic propaganda detection. Given a multigenre text which could be a news paragraph or a tweet, the objective is to identify the propaganda technique employed in the text along with the exact span(s) where each technique occurs. We approach this task as a sequence tagging task. We utilise AraBERT for sequence classification and implement data augmentation and random truncation methods to mitigate the class imbalance within the dataset. We demonstrate the importance of considering macro-F1 as well as micro-F1 when evaluating classifier performance in this scenario. Our system scored 0.12 micro F1 score and ranked fifth among the participants.

1 Introduction

As propaganda aims at promoting certain narratives to achieve specific purposes or advance certain agendas(Da San Martino et al., 2019), it can be used to manipulate people's opinions towards certain policies or ideologies. Therefore, developing systems to identify propaganda techniques in texts can help individuals and organisations combat bias, promote more accurate information and make informed decisions on what to believe. This is particularly important in today's polarized political and ideological climates, where false or misleading information can be used to manipulate public opinion.

Early research on propaganda detection has focused mainly on article and document level classification in English such as (Rashkin et al., 2017) where the authors built a corpus of news articles from different sources and labelled them broadly into four categories: Propaganda, hoax, trusted, and satire. Moving from the document level to a fine-grained analysis of texts,(Da San Martino et al., 2019) conducted a fine-grained analysis of texts by detecting fragments that have propaganda and their type. In their research, the authors derived 18 propaganda techniques from literature and annotated articles from propagandist and non-propagandist news outlets. Subsequently,(Da San Martino et al., 2020) formulated the SemEval-2020 Task 11 on Detection of Propaganda Techniques in News Articles, comprising two subtasks: span identification and propaganda technique classification. The aim of the current ArAIEval Shared Task is to extend this line of work to the detection of propaganda in Arabic text. For the unimodal (text) propagandistic technique detection task (Hasanain et al., 2024b), given a multigenre text which could be a news paragraph or a tweet, the objective is to identify the propaganda technique employed in the text along with the exact span(s) where each technique occurs[1].

Inspired by the success of BERT-based approaches to propaganda detection in English, we approach this task as a sequence tagging task and utilise an AraBERT(Antoun et al., 2020) token classification model. However, it is well known that class imbalance can be a problem for supervised machine learning methods. With few examples of minority classes in the training data, classifiers will tend to learn better representations of the majority classes. Class imbalance can be mitigated using undersampling, oversampling and other data augmentation techniques. However, a testing regime which uses micro F1 score over imbalanced classes will naturally favour methods which are biased towards predicting majority classes. In extreme cases, a classifier could fail to predict any class except the most frequently occurring class and appear to be better than a classifier which can detect less frequently occurring classes. We argue that the macro F1 score should also be considered as it provides a more balanced assessment of the model's classi-

[1] https://araieval.gitlab.io/task1/

Proceedings of The Second Arabic Natural Language Processing Conference, pages 524–529
August 16, 2024 ©2024 Association for Computational Linguistics

fication performance across all classes, regardless of their individual frequencies. Equipped with a classifier that can detect potential cases of minority classes, it would then be possible to bring a human into the loop to curate these examples and ultimately bootstrap a better classifier.

In this paper, we explore the effect that class imbalance within the ArAIEval dataset has on an off-the-shelf sequence tagger using AraBERT. We also investigate the effect of a data augmentation method and a simple random truncation method on models' performance using both the micro F1 score and the macro F1 score.

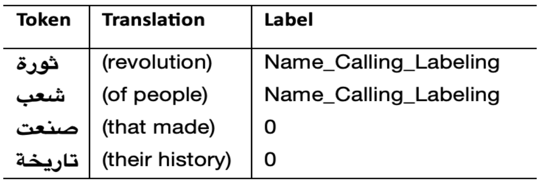
The rest of this paper is organised as follows: we discuss the dataset and the class distribution (Section 2); we outline the AraBERT token classification model used as well as the data augmentation and random truncation methods we have employed to mitigate class imbalance (Section 3); we provide a detailed discussion of the evaluation metrics (Section 4); we discuss our experimental results (Section 5);6 we provide an overview of propaganda detection related work; and finally present our conclusions and directions for future work (Section 7).

2 Data

In this paper, we are using the ArAIEval24 dataset[2]. Training, development, and test data files were made available in JSON format. Datasets with annotated instances of propaganda techniques and spans were provided for training and development. Each instance of a propaganda span is represented as a dictionary containing the start and end characters of the span in addition to the propaganda technique, and the corresponding text snippet.

2.1 Data Processing

We carried out several steps to process the ArAIEval24 data so it could be used for our AraBERT token classification model training and evaluation. In the datafiles, annotated propaganda spans were given as snippets without context so we aligned spans with their respective context sentences using the start and end information in the labels column to further process them. Next, tags marking the beginning and end of the propaganda span within the sentence were added to facilitate further processing. To accommodate cases where sentences have overlapping spans, each span was aligned to its respective context sentence individually, one at a time. The AraBERT token classification model was given the data in a token label format where each token within the tagged propaganda span was labelled according to its respective propaganda technique. In contrast, tokens outside the span were labelled as zero indicating that they are not propaganda tokens. Figure1 shows how labels are applied to tokens. An English translation is provided for each token to aid non-Arabic readers.



Token	Translation	Label
ثورة	(revolution)	Name_Calling_Labeling
شعب	(of people)	Name_Calling_Labeling
صنعت	(that made)	0
تاريخة	(their history)	0

Figure 1: Token Label Train Data

2.2 Class Distribution

Figure 2 shows the distribution of classes in the training dataset, where the size of each class is the number of spans which are labelled as belonging to that class. The two majority classes, loaded language and name-calling, constitute 55.72% and 14.23% of the dataset respectively whereas the minority classes including Appeal-to-Popularity, Red-Herring, Whataboutism, and Straw-Man etc. each constitute approximately 0.23% of the dataset.

3 Methodology

We approach the unimodal (text) propagandistic technique detection task as a sequence tagging task and we utilise AraBERT token classification model for this purpose. We fine-tuned the AraBERT model for token classification using the AutoModelForTokenClassification class from the Transformers library. The model was trained using the Trainer class with the following key parameters: an output directory AraBERT_token_classification for saving checkpoints and logs, evaluation and save strategies set to epoch, an initial learning rate of 2e-5, 10 training epochs, a weight decay of 0.01, and automatic pushing of the model to the Hugging Face Model Hub post-training.

3.1 Data Augmentation

To mitigate the class imbalance, we applied two different techniques. First, we applied data augmentation to increase the size of the minority classes with

[2] https://araieval.gitlab.io/task1/

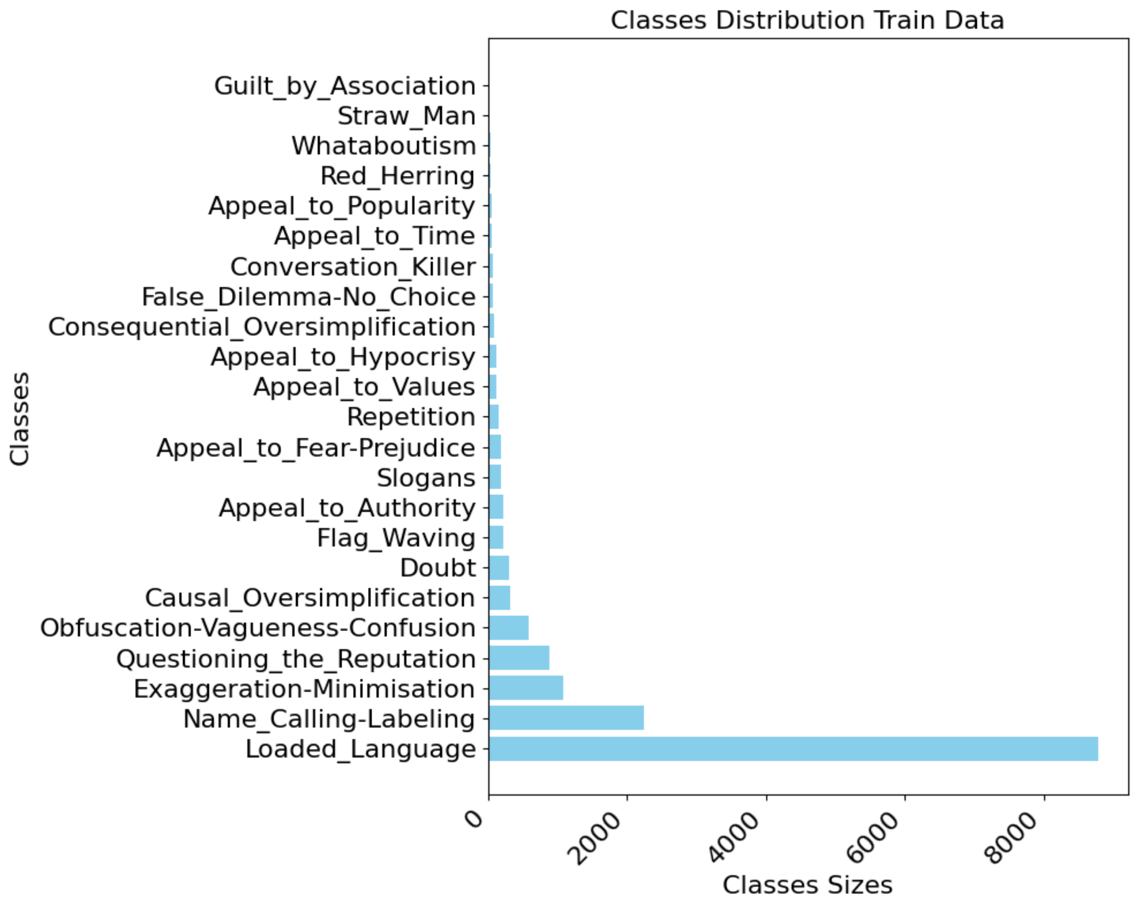


Figure 2: ArAIEval24 Train Dataset Class Distribution

a view to improving the overall performance of the model. Random masking was used on our propaganda sentences to generate the synthetic data for augmentation. Initially, we carried out some data pre-processing where we incorporated snippets into their context and snippets were marked by special tags <BOP> and <EOP> to mark their beginning and end within sentences. The whole sentence context was incorporated to make sure that the model had contextual data to improve prediction quality. Masking was applied as follows. If there were n words in the propaganda snippet, n different versions were created where one word was masked in turn and changed. After applying masks to our data, we ran the data through the AraBERT MLM model and used the fill mask function[3] from the Hugging Face library to make predictions. Based on preliminary experiments, a sample of the synthetic data was generated and all classes were incremented by 1000 instances to increase their size and representation in the training dataset. This led to a more balanced dataset with all classes having between 1000 and 10,000 training instances.

3.2 Random Truncation

Considering also the number of tokens which are labelled "0" (no-propaganda), the problem of class imbalance becomes further exacerbated as such tokens constitute approximately 89.05% of the total sum of tokens in the training corpus. Accordingly, we applied random truncation targeting tokens outside the propaganda spans to reduce the no-propaganda tokens' dominance within the dataset. The goal is to rebalance the distribution of classes and thus improve the model's performance. Specifically, we took each sequence containing a propaganda span and randomly truncated it to the left and the right of the propaganda span. Random truncation removed all of the no-propaganda tokens other than a random number between 0 and 5 on either side of the propaganda span. Random truncation significantly reduced the number of tokens for the (no-propaganda) class by 31.41%, where the total number of tokens before truncation was 586,242, representing approximately 89.05% of the

[3] https://huggingface.co/tasks/fill-mask

total sum. After truncation, this number decreased to 120,539, which accounted for approximately 64.88% of the total sum.

4 Evaluation Metrics

For our experiment results, we are reporting using the macro F1 score as well as the micro F1 score. Our use of the macro F1 score is motivated by the inherent class imbalance within the dataset. The micro F1 score computes a weighted average of F1 scores across all classes, taking into account the number of instances of each class in the dataset, and is essentially equivalent to computing accuracy (Da San Martino et al., 2020) In the case of an imbalanced dataset where one or two classes have much higher frequency than other classes, the micro F1 score tends to be dominated by the performance of the majority class(s). It can be a useful measure of overall performance but can also be misleading particularly with regard to a model's effectiveness in classifying minority classes. In contrast, the macro F1 score computes the F1 score for each class individually and then averages these scores across all classes, assigning equal weight to each class regardless of its size. This approach ensures that the evaluation metric is not biased towards the dominant class(s) and accurately captures the model's performance across all classes, including minority classes. In the context of imbalanced data representation, the macro F1 score provides a more balanced assessment of the model's classification performance.

5 Results and Discussion

Here we present our evaluation results on the test set provided. Figure 3 shows the F1 scores for all of the classes for the standard non-augmented AraBERT model, the augmented AraBERT model and the random truncation AraBERT model. Table 1 summarises the macro and micro F1 scores for the different models on the test dataset.

From Figure 3, we note that the standard non-augmented AraBERT model fails to predict more than 12 of the minority classes. Applying data augmentation has worsened the model's performance with the overall performance decreasing from a macro F1 score of 0.05 to 0.03. The decline in the model's performance following augmentation could be attributed to the existing class imbalance between propaganda and non-propaganda tokens. This imbalance possibly led to a disproportionate representation in the augmented data, thereby exacerbating the original class imbalance and consequently compromising the model's overall effectiveness.

In contrast to non-augmented and augmented AraBERT models, applying truncation has remarkably improved the overall performance of the AraBERT model from a macro F1 score of 0.05 to 0.10. Figure3 shows that the model performance has improved across the board and most of the minority classes are now being predicted.

However, just considering the micro-F1 scores in Table 1, we would not reach these conclusions. We would see that the micro-F1 score decreases using augmentation. Therefore, it is critical to consider other evaluation metrics such as macro-F1 if we wish to develop models which can predict more than the majority classes.

Model	Macro F1	Micro F1
AraBERT Basic	0.05	0.06
AraBERT Augmented	0.03	0.05
AarBERT Rand-Truncated	0.10	0.12

Table 1: Macro Vs Micro F1 scores

6 Related Work

Early research on propaganda detection has focused mainly on the article and document level classification (Rashkin et al., 2017);(Barrón-Cedeño et al., 2019). Moving from the document level,(Da San Martino et al., 2019) are the first to conduct a fine-grained analysis of texts detecting fragments with propaganda and identifying their type. In contrast to (Da San Martino et al., 2019) line of research which focuses on identifying propaganda in news articles,(Vijayaraghavan and Vosoughi, 2022) shift their attention to propaganda detection on social media, specifically Twitter. They introduced an end-to-end Transformer-based model enhanced with a multi-view approach that incorporates; context, relational data, and external knowledge into the representations. Additionally, (Alam et al., 2021) introduced an annotated dataset containing 950 Arabic further enriching the resources available for social media analysis. Recently,(Hasanain et al., 2024a) attempted leveraging GPT-4(OpenAI, 2023) to detect propaganda spans and identify propaganda techniques using a zero-shot setting. Furthermore, there have been initiatives such as (Dimitrov et al., 2021) and (Alam et al., 2024) to address multimodal content like

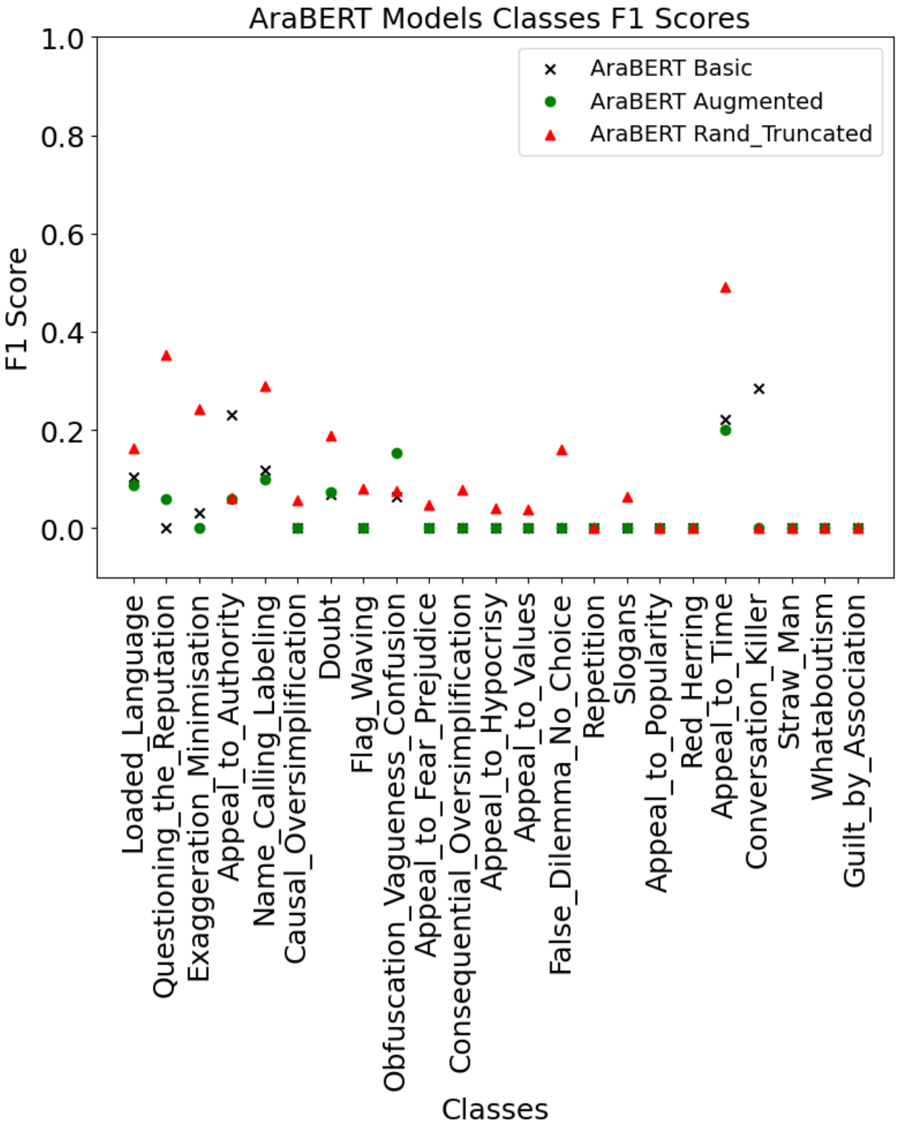


Figure 3: AraBERT Models Classes F1 Scores

memes, expanding the scope of propaganda analysis to include various media formats.

7 Conclusion and Future Work

In this paper, we introduced random truncation as a methodology to address the class imbalance in the ArAIEval24 train dataset. our method has proved to be effective in enhancing the model performance across the majority of the classes. Additionally, we emphasize the importance of employing the macro F1 score, instead of or as well as the micro F1 score, as a more suitable evaluation metric for the task of propaganda detection. Considering the inherent class imbalance within the dataset, the macro F1 score ensures a balanced evaluation of the models' performance compared to the micro F1 score which is susceptible to bias influenced by the dominant class(s) in an imbalanced dataset. In the future, we will explore applying active learning strategies as another method for discovering and sampling more instances representing minority classes aiming to enhance the class balance of the train data. Equipped with a more effective minority class classifier, we investigate applying it to large unannotated corpora, curating the examples found and ultimately bootstrapping a better classifier.

8 Acknowledgement

The researcher is sponsored by the Ministry of Higher Education of the Arab Republic of Egypt.

References

Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024. ArMeme: Propagandistic content in arabic memes. *arXiv preprint arXiv:2406.03916*.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

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 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024b. ArAIEval Shared Task: Propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok. Association for Computational Linguistics.

OpenAI. 2023. GPT-4: Technical report. Technical report, OpenAI.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States. Association for Computational Linguistics.