rematchka at ArAIEval Shared Task: Prefix-Tuning & Prompt-tuning for Improved Detection of Propaganda and Disinformation in Arabic Social Media Content

Reem Abdel-Salam
Cairo University, Faculty of Engineering, Computer Engineering / Giza, Egypt
reem.abdelsalam13@gmail.com

Abstract

The rise of propaganda and disinformation in the digital age has necessitated the development of effective detection methods to combat the spread of deceptive information. In this paper, we present our approach proposed for the ArAIEval shared task: propaganda and disinformation detection in Arabic text. Our system utilized different pre-trained BERT based models, that make use of prompt-learning based on knowledgeable expansion and prefix-tuning. The proposed approach secured third place in subtask-1A with a 0.7555 F1-micro score, and second place in subtask-1B with a 0.5658 F1-micro score. However, for subtask-2A & 2B, the proposed system achieved fourth place with an F1-micro score of 0.9040, and 0.8219 respectively. Our findings suggest that prompt-tuning-based & prefix-tuning based models performed better than conventional fine-tuning. Furthermore, using loss-aware class imbalance, improved performance.

1 Introduction

With the growing popularity of social media in our current society, platforms such as Twitter, and Reddit have become critical tools for influencing people. People on social media prefer to express their opinions, points of view more freely, and share information. However, these platforms can be used to deceive and manipulate individuals. In addition to spreading rumors, and fake news. This can be done through propaganda techniques. Propaganda refers to the systemic dissemination of information, ideas, or opinions, often through biased or misleading means, with the intention of influencing or manipulating public perception, attitudes, behaviors, or beliefs. It is a persuasive communication technique employed by individuals, organizations, or governments to shape public opinion and advance specific agendas. The rise of propaganda and disinformation has necessitated the development of effective detection methods to combat the spread of deceptive information. With the advent of pre-trained language models, there has been a significant advancement in the field of natural language processing (NLP), offering promising opportunities for combating the dissemination of false information. Several works have been proposed to improve the identification of persuasion techniques in text as the recent shared-task propaganda detection in Arabic (Alam et al., 2022). (Samir et al., 2022) and (Laskar et al., 2022) utilized AraBERT for this task. (Attieh and Hassan, 2022) utilized A multi-Task learning model, which includes a shared AraBERT encoder and task-specific binary classification layers. This model has been trained to learn one binary classification task per propaganda approach jointly. In this paper, we present our solution to the ArAIEval shared task (Hasanain et al., 2023). The ArAIEval shared task is held with the 1st Arabic Natural Language Processing Conference co-located with the EMNLP 2023. The goal of the task is to build models for identifying propaganda and disinformation in Arabic content. The shared task consists of two tasks. The first task is persuasion technique detection in Arabic text. The second task is disinformation detection in the text.

This paper describes the system developed for addressing propaganda and disinformation detection in text, for both subtasks. Given that a key challenge in this task is the unbalanced distribution of the dataset. Additionally, the contextual nature of language and the cultural nuances involved in the text. We follow best practices from recent work on enhancing model generalization and robustness, by using different parameter-efficient techniques (PEFT), contrastive loss, adversarial training, and loss-aware class imbalance methods. The Parameter-Efficient Fine Tuning (PEFT) is a technique used to improve the performance of pre-trained language models on specific downstream tasks. PEFT methods freeze the pretrained model

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 536–542

parameters during fine-tuning and put a few trainable parameters (the adapters) on top of it. The adapters are taught how to pick up knowledge appropriate to a given task. PEFT of pre-trained language models has recently demonstrated remarkable results, effectively matching the performance of full fine-tuning while utilizing significantly fewer trainable parameters (Fu et al., 2023; Liu et al., 2022; Wang et al., 2022), thereby addressing storage and communication constraints. Such approaches include prefix-tuning (Li and Liang, 2021), prompt-tuning (Hu et al., 2021b), soft-prompting (Lester et al., 2021) and LoRa (Hu et al., 2021a). Adversarial training (AT) (Goodfellow et al., 2014) is a method to improve the model's resistance to adversarial examples and acts as a regularizer. The key is to disturb the input example using a gradient-based perturbation, and then train the model on both clean and perturbed examples. Contrastive loss is one of the first training objectives that was used for contrastive learning. It takes as input a pair of samples that are either similar or dissimilar, and it brings similar samples closer and dissimilar samples far apart in embedding space (Khosla et al., 2020). Such loss has shown model performance improvement compared to cross-entropy on multiple problems (Chi et al., 2022; Chen et al., 2022; Pan et al., 2022).

The rest of the paper goes as follows: section 2 gives an overview of the dataset, section 3 discusses the proposed methods, section 4 shows experimental results, and section 5 concludes the paper.

2 Data

The dataset used has been provided by the organizers for the ArAIEval shared task. Table 1 summarizes the distribution of the provided dataset. For subtask-1A the dataset consists of the text of Arabic tweets, the type of the text whether it is a tweet or text, and the label. The train, validation, and consist of 2427, and 259 examples. The provided data is unbalanced as for the non-persuasion class 509 is presented whilst, the other class 1918 example is presented. For subtask-2A&2B the provided dataset consists of the text and the label. In subtask-2A, the distribution of labels in the train-set goes as follows: 2656 examples for the disinformation class, and 11491 examples for the non-disinformation text class. The distribution of labels in subtask-2B in the train-set is as follows: hate speech 1512 examples, 453 examples for the spam class, 500 examples for the offensive class, and 191 examples for the rumor class. Accordingly, a major issue in this dataset is the nature of the unbalance of the class distribution, which poses a challenge.

Task	Train-size	Dev-size	Test-size
Subtask-1A	2427	259	503
Subtask-1B	2427	259	503
Subtask-2A	14147	2111	3729
Subtask-2B	2656	397	876

Table 1: Distribution of the provided dataset

3 Methodology

This section presents the various approaches used while developing the final models: a weighted ensemble of BERT-based models.

3.1 Task-1

Task-1 was composed of two subtasks, subtask-1A and subtask-1B. The goal of subtask-1A is to detect whether a given text contains content with a persuasion technique. The goal of subtask-1B is to identify which of the 24 propaganda techniques is used in a given text. In order to address these subtasks, we tried a variety of ways. The majority of the models employed were BERT-based, such as MARBERT (Abdul-Mageed et al., 2020) and AraBERT (Antoun et al., 2020).

subtask-1A In subtask-1A two methods were used: conventional fine-tuning and prefix-tuning. In order to make the model more robust so that similar inputs derive semantically similar outcomes two approaches were explored fast gradient methods (FGM) (Wang et al., 2021) and supervised contrastive learning (Chen et al., 2022). In addition, back-translation between Arabic and English languages was used as an augmentation, to upsample the dataset for the lower class. Prefix tuning is an additive technique that only attaches a continuous set of task-specific vectors to the input's beginning. In each layer of the model, the hidden states are only added and the prefix parameters are optimised. The input sequence's tokens can still serve as virtual tokens to the prefix. Fast Gradient Method (FGM), is a popular technique for generating adversarial examples. It works by adding small, carefully crafted perturbations to the input data, in our case, the perturbations are added to the model's embedding, such that the model's prediction changes to a

wrong answer. The Fast Gradient Method is based on the concept of a "fast gradient" - a gradient that is calculated with respect to the input data, instead of the model's parameters.

subtask-1B The challenge of this subtask was to correctly identify labels for each text, in a given unbalanced dataset. To address these issues two approaches have been investigated: 1) loss aware class imbalance such as Asymmetric loss for multi-label classification (Ridnik et al., 2021), and Distribution Balanced Loss (Wu et al., 2020) 2) balanced data-Sampler for multi-label problems. In this task all models were trained using prefix-tuning.

3.2 Task-2

Task-2 was composed of two subtasks, subtask-2A and subtask-2B. The goal in subtask-2A is to classify whether a given text is disinformation or not. However, in subtask-2B the goal was to predict the disinformation class of a given text.

subtask-2A In this subtask, the same experiments conducted in subtask-1A were used in this subtask.

subtask-2B In this subtask, prompt-tuning was utilized using openprompt library (Ding et al., 2021). Prompt tuning is the process of feeding front-end prompts into the model in the context of a specific task. These prompts could be either text related to the task or virtual tokens. Prompt tuning is used to guide a model toward a particular prediction. Prompts are only introduced into the input embedding sequence and this embedding is fed to the language model head and output to the linear classification head, as shown in the figure 1. One of the difficulties in promoting is the design of the prompt and the model's output. For the prompt, we used [MASK] فئة المعلومات المضللة ("The disinformation class is [MASK]"), and For the output, we have used label names translated into Arabic. Two models were used: AraBERT and AraGPT.

Experimental Set-up for the fine-tuned models the learning rate was set to 4e-5 or 4e-6, a cosine-annealing learning rate scheduler was used, the model's weight decay was set to 1e-8 and the length of the sentence for tokenization was set to 128 or 256. During training, batch size was set to 32, and at the end of each epoch, the model was evaluated on dev-set. The best-performing model in terms of F1-micro is saved.

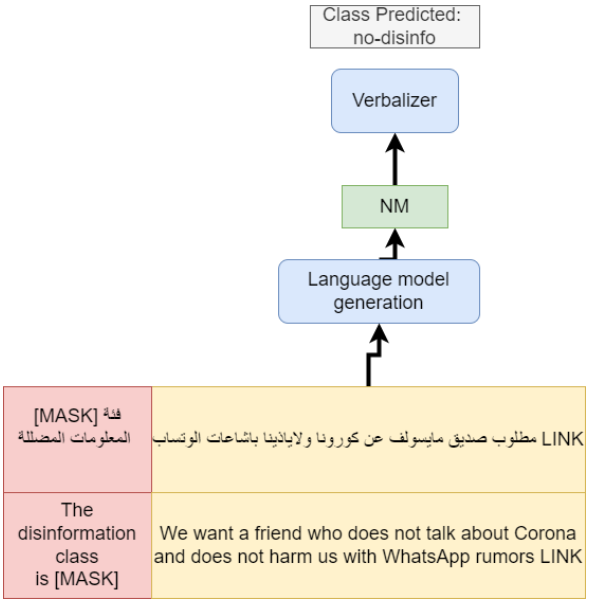

Figure 1: Prompt-tuning architecture.

4 Results and Discussion

In this section, The performance of the model is reported based on the official metric during dev-phase and test-phase. The official metric used for all tasks is the micro average F1-score. Table 3 shows results for subtask-1A on dev-set and test-set. In the dev-set, the outperforming model was Arabert v2 with prefix-tuning, which comes in second place with MARBERT with prefix-tuning and contrastive learning. Surprisingly, the performance of the model is switched in the test-set. It is noticed that high performance in dev-set does not necessarily mirror the test-set. The reason behind it is the nature of the training. For instance, contrastive loss and FGM make models robust so that similar inputs derive in semantically similar outcomes. Table 4 shows results in subtask-1B. It could be concluded that class-aware loss function with a balanced sampler improves model performance over simple binary cross-entropy loss with random samplers. Table 9 and 2 show results in subtask-2A and subtask-2B, similar to subtask-1A outperforming models in dev-set are interleaved in test-set. Tables 5,6,7, and 6 shows different teams run in the shared task. For single models, in table 3 both Arabert with Prefix-tuning and MARBERT with Prefix-tuning contrastive loss with Cross entropy loss show high competence with submitted models on the leaderboard 5, as they could have secured first and second places. For subtask-1B based on tables 6,4, the ensemble model seems to be on par

with single models. Since, Arabert with Asymmetric Loss for a single model run, shows similar results to the ensemble and would have secured the same place in leaderboard. For subtask-2A&2B based on tables 9,7, 2 and 8, the ensemble model seems to be the best solution over single models. Non of the single models could have secured a higher place than the ensemble model.

4.1 Error Analysis

Further investigations have been carried out to analyze the potential limitations of the system. For subtask-1A, the model could not correctly identify the following text into the correct class: persuasion class.

شهدت مجموعة من مدن المملكة، اليوم الجمعة (١٦ أكتوبر)، أول صلاة جمعة في زمن كورونا، بعد أزيد من ٧ شهور من تعليقها، من طرف السلطات للحد من تفشي فيروس كورونا المستجد.

Today, Friday (October 16), a group of cities in the Kingdom witnessed the first Friday prayer in the time of Corona, more than 7 months after it was suspended by the authorities to limit the spread of the new Coronavirus. The reason behind this is that the model has no knowledge of previous information about coronavirus lockdown, and its consequences. Therefore, it is hard to assess the facts in the text. Another miss-classification error, where true class is non-persuasion is

عترف بأن حزناً عميقاً راح يعبر القلب لحظة وقوع ذلك الطائر المهاجر الذي كان يقصد تلك الفيافي ليرتاح على حصاها قليلاً ثم لا يلبث أن يغادر المكان الى حيث الدفء الذي يبحث عنه.. فلم يجد الا الغدر وخيانتي للضيف.

He admitted that a deep sadness began to cross his heart the moment that migratory bird fell, which was heading to that desert to rest on its pebbles for a little while, and then quickly left the place for the warmth he was looking for.. He found nothing but treachery and my betrayal of the guest. The model failed to understand that the provided text is a poem rather than a piece of news. So it could be concluded that some of the errors are related are due to the model not able to handle different domains and gain knowledge about them and their differences.

Figure 2, shows model MARBERT performance in subtask-2B. The model confuses between hate speech class and the offensive class. As well as, between the offensive class and the rumor class.

Model	F1-micro Dev-set	F1-micro test set
Arabert v2	78	81
Aragpt	79	77
Final Model	-	82.19

Table 2: Results on our dev-set and test-set for the developed models in subtask-2b

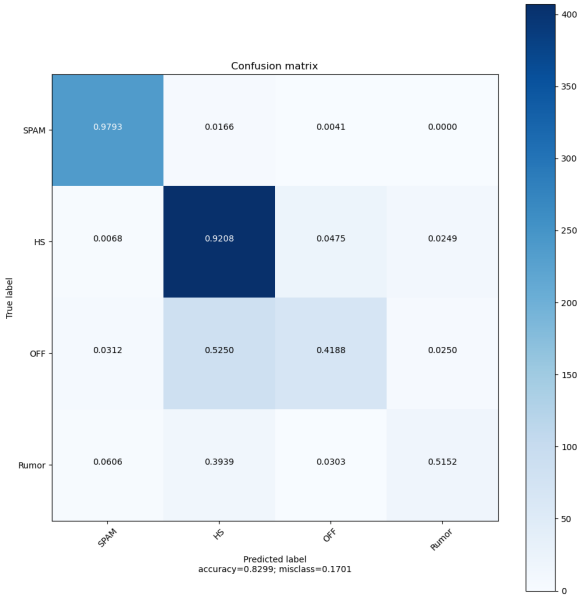

Figure 2: Confusion matrix of the predictions of the submission-3 model in subtask 1 on the dev-set.

5 Conclusion

In this paper, the results and the main findings of ArAIEval shared task were presented, in which different experiments were carried out with MARBERT, Arabert v2, and Aragpt models. Our models secured third place in subtask-1A, second place in subtask-1B, and Fourth place in subtask-2A&2B. Our proposed solution is an ensemble of different BERT-based models. These Models are developed differently, some are trained using prefix-tuning, and others are trained using fine-tuning and prompt-tuning. leverages fine-tuned, per-trained models. In addition, training tricks were utilized as FGM, contrastive learning, and balanced sampler. In future efforts, we plan to further improve our model to better handle data-imbalance constraints and world knowledge needed to improve model performance.

Model	Technique	F1-micro Dev-set	F1-micro test set
Arabert v2	Prefix-tuning	84.8	75.8
	Prefix-tuning Back Translation	77	72.7
	Prefix-tuning FGM	85	74.2
	Prefix-tuning Type of text specificed	85.9	73.9
	Prefix-tuning Focal loss	83.8	75
	Fine-tuning	83.6	72.2
MARBERT	Prefix-tuning contrastive loss with Cross entropy loss	84.5	76.5
Final Model	Ensemble	-	75.55

Table 3: Results on our dev-set and test-set for the developed models in subtask-1A

Model	Technique	F1-micro Dev-set	F1-micro Test-set
Arabert V2	Resample Loss Sampler	66	54
	Binary Cross Entropy Loss	62	51
	Asymmetric Loss Sampler	64.89	56
Final Model	Ensemble	-	56.58

Table 4: Results on our dev-set and test-set for the developed models in subtask-1B

Team	Micro F1
HTE	76.34
KnowTellConvince	75.75
rematchka	75.55
UL & UM6P	75.15

Table 5: Leaderboard results on test-set for subtask-1A

Team	Micro F1
UL&UM6P	56.66
rematchka	56.58
AAST-NLP	55.22
Itri Amigos	55.06

Table 6: Leaderboard results on test-set for subtask-1B

Team	Micro F1
DetectiveRedasers	90.48
AAST-NLP	90.43
UL&UM6P	90.40
rematchka	90.40
PD-AR	90.21

Table 7: Leaderboard results on test-set for subtask-2A

Team	Micro F1
DetectiveRedasers	8356
UL&UM6P	83.33
AAST-NLP	82.53
rematchka	82.19
superMario	8.208

Table 8: Leaderboard results on test-set for subtask-2B

Model	Technique	F1-micro Dev-set	F1-micro test set
Arabert v2	Prefix-tuning	88.3	89.2
Arabert v2	Prefix-tuning Back Translation	90.01	89.5
Arabert v2	Prefix-tuning FGM	89.8	88
Final Model	Ensemble	-	90.40

Table 9: Results on our dev-set and test-set for the developed models in subtask-2A

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, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). 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*.

Joseph Attieh and Fadi Hassan. 2022. Pythoneers at wanlp 2022 shared task: Monolingual arabert for arabic propaganda detection and span extraction. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 534–540.

Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. 2022. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv preprint arXiv:2201.08702*.

Jianfeng Chi, William Shand, Yaodong Yu, Kai-Wei Chang, Han Zhao, and Yuan Tian. 2022. Conditional supervised contrastive learning for fair text classification. *arXiv preprint arXiv:2205.11485*.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2023. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12799–12807.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat. 2023. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2021b. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Sahinur Rahman Laskar, Rahul Singh, Abdullah Faiz Ur Rahman Khilji, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022. Cnlp-nits-pp at wanlp 2022 shared task: Propaganda detection in arabic using data augmentation and arabert pre-trained model. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 541–544.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91.

Ahmed Samir, Abu Bakr Soliman, Mohamed Ibrahim, Laila Hesham, and Samhaa R El-Beltagy. 2022. Ngu_cnlp at wanlp 2022 shared task: Propaganda detection in arabic. *WANLP 2022*, page 545.

Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. 2021. Adversarial training with fast gradient projection method against synonym substitution based text attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13997–14005.

Yaqing Wang, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022. Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models. *arXiv preprint arXiv:2205.12410*, 1(2):4.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer.