AraDetector at ArAIEval Shared Task: An Ensemble of Arabic-specific pre-trained BERT and GPT-4 for Arabic Disinformation Detection

Ahmed Bahaaulddin
Middle Technical University
ahmedbahaaulddin@mtu.edu.iq

Vian Sabeeh
Middle Technical University
viantalal@mtu.edu.iq

Hanan M. Belhaj
The Libyan Academy
h.belhaj@it.lam.edu.ly

Serry Sibaee
Serrytowork@gmail.com

Samar Ahmad
Samar.sass6@gmail.com

Ibrahim Khurfan
ibraheemkhurfan@gmail.com

Abdullah I. Alharbi
King Abdulaziz University
aamalharbe@kau.edu.sa

Abstract

The rapid proliferation of disinformation through social media has become one of the most dangerous means to deceive and influence people's thoughts, viewpoints, or behaviors due to social media's facilities, such as rapid access, lower cost, and ease of use. Disinformation can spread through social media in different ways, such as fake news stories, doctored images or videos, deceptive data, and even conspiracy theories, thus making detecting disinformation challenging. This paper is a part of participation in the ArAIEval competition that relate to disinformation detection. This work evaluated four models: MARBERT, the proposed ensemble model, and two tests over GPT-4 (zero-shot and Few-shot). GPT-4 achieved micro-F1 79.01% while the ensemble method obtained 76.83%. Despite no improvement in the micro-F1 score on the dev dataset using the ensemble approach, we still used it for the test dataset predictions. We believed that merging different classifiers might enhance the system's prediction accuracy.

1 Introduction

Approximately 66% [1]of individuals in the Middle East utilize social media to seek out daily news. The rise of rapid development in social media and online communication, such as chat platforms (WhatsApp, Facebook Messenger, Snapchat, and LINE), have emerged as prevalent means to facilitate the widespread dissemination of disinformation at an unprecedented pace. Disinformation is the phenomenon that refers to how individuals or groups can be deceived or manipulated by false or misleading information. As disinformation spreads gradually, it can boost existing biases, polarize viewpoints, and hinder constructive dialogue, compromising the collaborative spirit essential for a healthy democracy. This far-reaching phenomenon can affect opinion decision-making and can threaten different foundations of democratic societies by eroding public trust in different institutions and planting seeds of divisions among communities (Himdi et al., 2022; Shu et al., 2020; Freelon and Wells, 2020). To detect disinformation and prevent it from spreading, modern methods use transformer-based architectures that are trained specifically on Arabic text and are available in public, such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021).

This paper outlines our participation in the disinformation detection (Task 2) of the ArAIEval Shared Task (Hasanain et al., 2023). We introduce a method incorporating three distinct classifiers: the MARBERT Pre-trained Language Model (PLM) and both zero-shot and few-shot models. Our objective is to improve the accuracy of disinformation identification in tweets by adopting a majority voting ensemble strategy. The subsequent sections are structured as follows: Section 2 reviews prior studies; Section 3 describes our proposed method; Section 4 details our experimental result; and finally, we conclude with a summarisation of our main findings.

2 Related Work

Nowadays, various types of disinformation have swiftly disseminated across social media platforms and digital news outlets, Each possessing distinct attributes and objectives to deceive and influence people. Due to the simplicity of sharing data online, it has become challenging to differentiate between trustworthy information and fake ones (Aïmeur et al., 2023). Many research studies have been con-

[1] https://www.mideastmedia.org/survey/2017/chapter/social-media/#s225

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 530–535

ducted to detect disinformation. In this section, we provide a concise overview of the recent research on disinformation. In Their Study ABOUT DISINFORMATION DETECTION, Bahurmuz N. et al. 2017 used two transformers, AraBERT and MARABERT. The proposed paradigm removes all the non-textual, non-linguistic URL features to get a real dataset. Two sampling techniques have been used to solve the unbalanced dataset. Transformers models were trained by fine-tuning the hyper-parameters using freezing model techniques. MARABERT shows better performance results (Bahurmuz et al., 2022). In 2021, Al-Yahya M et al. examined various neural networks and transformer models for Arabic fake news detection. The experiments were conducted by using document and word embedding to test multiple neural network models like CNN, RNN, and GRU, then compared to the transformers like (ARABERT V1, ARABERT V2, ArElectra, QARiB, Arbert, and Marbert. QARiB obtained high accuracy scores compared to the limitation of small data size, repeated tweets, and noisy tweets that do not belong to any class (Al-Yahya et al., 2021).

ALbalawi. R. et al. 2022 proposed a model that relies on textual visual features to detect disinformation. They used MRABERT for text feature extraction, while RESNET-50 was used to extract image features. The text and visual features were combined and input into one multi-modal classifier to detect rumors from non. Early fusion of features achieved an accuracy of 0.85. The efficiency of the proposed model could not outperform the text-based models in accuracy. This is due to the size of the dataset (R. M. Albalawi et al., 2023).

Obeidat R. et al. (2022) worked on collecting a dataset related to COVID-19 disinformation news from Twitter; this dataset was the first Arabic COVID-19 dataset comprising about 6.7K tweets. Word cloud has been used to obtain crucial words to analyze both real and fake news. They also prepared a version of ARaBert trained based on COVID-19 tweets known as AraBERT-COV19. To reach a more accurate result than previous models, authors have been dependent on preparing and labeling the collected dataset manually. (Obeidat et al., 2022).

Hate speech and fake news can work together as a powerful weapon against society; for example, an article claiming that a particular group of people is planning to commit violence can justify hate speech against that group. This can lead to real-world violence, as seen in cases such as the Rohingya genocide in Myanmar (Doncel-Martín et al., 2023). A study by researchers at the University of Southern California's Information Sciences Institute found that 20 percent of tweets containing hate speech were also fake (Zheng et al., 2020).; Therefore, Ameur M. et al. (2021) used fine-tuned two pre-trained models, AraBERT COV19" and "mBERT COV19. The work aimed to build a model that can detect fake news about COVID-19 and hate speech simultaneously (Ameur and Aliane, 2021).

3 Methodology

The proposed system is composed of three distinct models. Raw tweets were prepared and pre-processed as inputs to the models, as outlined in Section 3.1. Sections 3.2, 3.3, and 3.4 explain the three models incorporated using an ensemble technique, as clarified in Section 3.4.

3.1 Preprocessing

Pre-processing was conducted using a methodology previously employed by various researchers (Duwairi and El-Orfali, 2014; Abu Farha and Magdy, 2019). The initial step involved eliminating unfamiliar symbols and characters, such as letters from different languages, punctuation, and diacritics. Emojis were retained because they may be used to express hate, obscenity, and abusive content (Mubarak et al., 2023). Additionally, certain letters that exhibited diverse forms within the original tweets were standardized to a singular form. For instance, characters like 'hamza' {ٱ,إ,أ} were substituted with {ا}, and the 't marabout {ة} was changed to {ه}.

3.2 Fine-tuning pre-trained Language Models

Due to the contextual nature of disinformation textual content, contextualized language models would be beneficial in addressing this task. Transformer architectures like BERT (Devlin et al., 2019) have demonstrated exceptional success across diverse NLP tasks. Our study employed three Arabic language models that have attained cutting-edge performance in various Arabic NLP applications. These models were fine-tuned for disinformation detection, enabling us to conduct a comparative analysis of their capabilities. The specific models employed are as follows:

AraBERT: Antoun et al. (2020) introduced a BERT-based model explicitly trained for the Arabic language. It emerged as the first Arabic-specific BERT model to achieve competitive results across most Arabic NLP tasks. This model was pre-trained on an extensive dataset encompassing 24 GB of text sourced from Wikipedia and various news outlets across the Arab region.

MARBERT: As presented by Abdul-Mageed et al. (2021), this model was designed for transfer learning in Arabic dialects. MARBERT's pre-training involved a massive dataset comprising 6 billion tweets, leading to state-of-the-art performance across multiple Arabic-language NLP tasks.

QARiB: Developed by Abdelali et al. (2021), this model underwent training using a mix of Modern Standard Arabic (MSA) and dialectal sources. The training dataset encompassed approximately 420 million tweets and 180 million sentences from news articles. Notably, the utilization of this combination of sources, comprising MSA and dialectal content for language model pre-training, is observed to enhance performance in classification tasks, according to the author's observations.

3.3 Large Language Models (LLMs)

Large Language Models (LLMs) have recently become essential in Natural Language Processing (NLP). They effectively utilize vast knowledge sources and deeply comprehend complex language details (Alyafeai et al., 2023; Zhang et al., 2023). One significant model in this area is OpenAI's GPT-4, an advanced language model supported by a transformer architecture and a massive 1.76 trillion parameters (OpenAI, 2023). While its effectiveness can vary by task, its strengths in sentiment analysis and emotion detection highlight its utility (Wang et al., 2023). Thus, the exploration of such models is essential for other NLP studies.

Zero-shot: We use GPT 4 as a zero-shot classifier; the model was never trained explicitly on our task. The key to our approach lies in the prompting strategy. Constructing effective prompts is vital; it is an implicit instruction to guide the model to understand and perform the desired classification, ensuring accurate and reliable outputs. Figure 1 illustrates an example of a zero-shot prompt, highlighting instructions for data and categorization. Considering GPT-4's potential, we designate its role as an "annotator expert." We introduce labels to steer the LLMs alongside the primary directive. The guidance specifies the format of the LLMs' responses, seeking to make any other adjustments.

Few-shot: The foundational research from Brown et al. (2020) highlighted the enhanced outcomes of few-shot learning relative to zero-shot configurations. In our work, we used a few-shot setting leveraging GPT-4. We selected nine examples from the available training data rather than selecting samples at random. To achieve this, we used the "sentence-transformers" library to obtain embeddings for Arabic tweets. It starts by choosing a random tweet from the training set and then iteratively picks the most dissimilar tweet based on cosine distance from the already selected ones. Specifically, each addition computes the sum of lengths to all previously selected tweets, ensuring a diverse selection. The "distiluse-base-multilingual-cased-v2" model is used for multilingual support, including Arabic. Since the proportion of the disinformation class in the training data set is small, we chose to increase the number of disinformation tweets (6 examples) compared to (3 models) for the class that does not contain misleading information. For each category, we applied the dissimilarity above selected samples approach. Figure 3 illustrates the details of the utilized prompt.

3.4 Ensemble

At this step, we have three individual classifiers: the best-performing Pre-trained Language Model (PLM) MARBERT, zero-shot, and few-shot models. Each model's output is a determination of whether a tweet is disinformation or not. By using different classifiers together, we can reduce their individual weaknesses and benefit from their strengths. Using an ensemble method, we employed a majority voting approach to merge the classifiers. We assume that combining multiple classifiers might generalize predictions on unseen tweets. This is based on the idea that multiple models may capture different aspects or features of the tweet, leading to a more comprehensive and reliable decision when combined.

4 Result

Given the nature of the shared tasks, we conducted our initial experiments on the development dataset and accordingly selected the best-performing method for delivering predictions on the test dataset. The organizers of this shared task have shared an annotated dataset sourced from

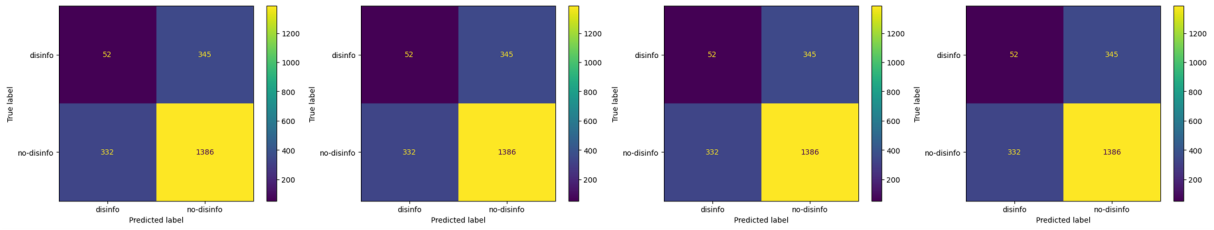

Figure 1: The confusion matrices for the three voter models: a) MARBERT, b) few-shot, c) zero-shot and d) our proposed method.

Pre-processing	AraBERT	QARiB	MARBERT
No	68.16%	69.04%	68.27%
Yes	68.64%	69.22%	**69.79%**

Table 1: Performance results (micro-F1) for fine-tuning three PLMs on the Disinformation Detection task.

Twitter (Hasanain et al., 2023). This dataset is noteworthy for being one of the most extensive publicly accessible Arabic datasets focused on disinformation content [2]. It contains 14126 tweets as a training set and 2115 tweets as a development set. Specifically, we compared three pre-trained models as well as the GPT models in Zero-Shot and View-Shot settings. We used the official evaluation measure adopted by the organizers (micro-F1).

We performed experiments to assess three PLMs trained specifically for Arabic: AraBERT, QARiB, and MARBERT. Each model was fine-tuned on the provided training set, and their performances were measured on the dev dataset using the official metric (micro-F1). Table 1 presents the performance results for fine-tuning three PLMs on the dev dataset. MARBERT showed the highest performance, securing a micro-F1 of 0.698, while QARiB was a close second at 0.693. When it comes to examining the pre-processing impact, the performance of models with preprocessing is better than without, with varying effects. MARABERT Results improved relatively with the use of preprocessing (1.52%), followed by an improvement of (0.68%), compared to a slight improvement of (0.18%) for QARiB model.

Additionally, we used two experimental settings: zero-shot and few-shot prompting strategies. Due to the cost of using such models, we used the pre-processed text based on previous experiments' findings. In future work, we will study the impact of pre-processed text for GPT models on Arabic user-generated text extracted from social media. Table 2 presents the performance of zero-shot and few-shot classifiers and our proposed ensemble approach. We observe that the performance of the zero-shot setup is generally higher than the few-shot setting, with a significant improvement of 15% (micro-F1). However, we found that the few-shot setting excelled when it came specifically to the disinformation class, as it predicted 337 tweets out of 397, while the zero-shot setting only recognized 168 tweets. This shows the importance of providing generalized examples, as we explained in Section 3.3. In future work, we will study the effect of the number of examples in general and their proportion for each class.

Finally, after studying and analyzing the performance of the models, we decided to take advantage of each one of them using a majority voting approach to merge three classifiers: MARBERT, few-shot and zero-shot prompting strategies. Although the micro-F1 score on the dev dataset was not improved using the ensemble approach, we used it to deliver predictions of the test dataset. We hypothesized that the system's ability to generalize by combining different classifiers may balance out better classification prediction. Table 2 presents the performance of our proposed ensemble approach on the test set, which is the official result for our participant. Figure 1 presents the confusion matrices for the three voter models and our proposed method.

5 Conclusion

Disinformation on social media may be biased in the society's collective opinion. Consequently, this may lead to social abuse action. Accordingly, social media needs an apparatus to help people reveal false claims. This study used three Arabic transformers for comparison (AraBERT, MARBERT, QARIB). From the experiments, we conclude that

[2] https://gitlab.com/araieval/wanlp2023_araieval/-/tree/main/task2

Model	macro-F1 disinfo	macro-F1	micro-F1
Dev Dataset			
MARBET	15.13%	48.84%	69.79%
Few-shot	41.86%	53.07%	55.74%
Zero-shot	43.08%	65.10%	79.01%
Ensemble	43.42%	64.43%	76.83%
Test Dataset - Formal submersion			
Ensemble	-	64.98%	74.87%

Table 2: Performance results for the three voter models: (MARBERT,few-shot and zero-shot) and our proposed method on the dev and test dataset.

there is an influence of pre-processing on model performance. To reach a generalized approach, two settings for the test were conducted depending on GPT-4: few-shot and zero-shot and one proposed ensemble learning. Zero-shot by GPT-4 achieves the best performance. Even though the ensemble approach did not boost the micro-F1 score on the dev dataset, we employed it in the test dataset, assuming that integrating various classifiers might improve prediction accuracy.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.

Maha Al-Yahya, Hend Al-Khalifa, Heyam Al-Baity, Duaa AlSaeed, and Amr Essam. 2021. Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches. *Complexity*, 2021:5516945. Publisher: Hindawi.

Zaid Alyafeai, Maged S Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. 2023. Taqyim: Evaluating arabic nlp tasks using chatgpt models. *arXiv preprint arXiv:2306.16322*.

Mohamed Seghir Hadj Ameur and Hassina Aliane. 2021. AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News and Hate Speech Detection Dataset. ArXiv:2105.03143 [cs].

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, Marseille, France. European Language Resource Association.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Naelah O Bahurmuz, Ghada A Amoudi, Fatmah A Baothman, Amani T Jamal, Hanan S Alghamdi, and Areej M Alhothali. 2022. Arabic Rumor Detection Using Contextual Deep Bidirectional Language Modeling. *IEEE Access*, 10:114907–114918. Publisher: IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis. Association for Computational Linguistics.

Israel Doncel-Martín, Daniel Catalan-Matamoros, and Carlos Elías. 2023. Corporate social responsibility and public diplomacy as formulas to reduce hate speech on social media in the fake news era. *Corporate Communications: An International Journal*, 28(2):340–352. Publisher: Emerald Publishing Limited.

Rehab Duwairi and Mahmoud El-Orfali. 2014. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science*, 40(4):501–513.

Deen Freelon and Chris Wells. 2020. Disinformation as political communication.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abdelhakim 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.

Hanen Himdi, George Weir, Fatmah Assiri, and Hassanin Al-Barhamtoshy. 2022. Arabic fake news detection based on textual analysis. *Arabian Journal for Science and Engineering*, 47(8):10453–10469.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2023. Emojis as anchors to detect arabic offensive language and hate speech. *Natural Language Engineering*, page 1–22.

Rasha Obeidat, Maram Gharaibeh, Malak Abdullah, and Yara Alharahsheh. 2022. Multi-label multi-class COVID-19 Arabic Twitter dataset with fine-grained misinformation and situational information annotations. *PeerJ Computer Science*, 8:e1151.

OpenAI. 2023. Gpt-4 technical report.

R. M. Albalawi, A. T. Jamal, A. O. Khadidos, and A. M. Alhothali. 2023. Multimodal Arabic Rumors Detection. *IEEE Access*, 11:9716–9730.

Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. 2020. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1385.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Han Zheng, Jinhui Li, Charles T. Salmon, and Yin-Leng Theng. 2020. The effects of exergames on emotional well-being of older adults. *Computers in Human Behavior*, 110:106383.