TAO at StanceEval2024 Shared Task: Arabic Stance Detection using AraBERT

Anas Melhem, Osama Hamed*, Thaer Sammar
Palestine Technical University - Kadoorie
(a.melhem,osama.hamed,thaer.sammar)@ptuk.edu.ps
*Corresponding author

Abstract

In this paper, we present a high-performing model for Arabic stance detection on the StanceEval2024 shared task part of ArabicNLP2024. Our model leverages AraBERTv1; a pre-trained Arabic language model, within a single-task learning framework. We fine-tuned the model on stance detection data for three specific topics: COVID-19 vaccine, digital transformation, and women empowerment, extracted from the Mawqif corpus. In terms of performance, our model achieves 73.30 *macro-F1* score for women empowerment, 70.51 for digital transformation, and 64.55 for COVID-19 vaccine detection.

1 Introduction

Stance detection is a critical domain within Natural Language Processing (NLP) that focuses on the automatic identification of attitudes or opinions expressed toward specific entities or propositions within textual data. With the ubiquity of the internet in our everyday lives, there has been a substantial increase in user-generated content online, where stances are primarily articulated through written comments or texts on forums, blogs, and social media platforms such as YouTube, Instagram, Facebook, and Twitter.

While significant research has yielded powerful stance detection models for languages like English, Arabic stance detection research still requires attention. Although Arabic language is one of the top ten global languages with over 420 million native speakers (Guellil et al., 2021; Cheng et al., 2021; Qu et al., 2024), Arabic presents unique challenges for NLP applications due to its rich and complex morphology and the variety of its dialects (Badaro et al., 2019).

Unlike formal media that utilize Modern Standard Arabic (MSA), social media platforms primarily feature dialectal Arabic with an informal writing style, often replete with spelling errors, abbreviations, irregular grammar, emojis, and symbols. This inherent disparity creates a significant obstacle to automated stance detection on social media.

Mawaqif dataset (Alturayeif et al., 2022) is considered the first Arabic dataset for stance detection. This dataset, which forms the foundation of the StanceEval () shared task, stands out for its extensive reach, encompassing more than 4000 tweets in multi-dialectal Arabic targeting three topics: "COVID-19 vaccine", "digital transformation", and "women empowerment". These are carefully annotated with stance, sentiment, and sarcasm polarities. The shared task () highlights two principal stance detection challenges:

(i) Stance detection through single-task learning-based models (STL): These models depend only on the stance data for model development and training.

(ii) Stance detection through multi-task learning-based models (MTL): These models leverage information, such as the sentiment and sarcasm of each tweet in order to boost the performance of the stance detection system.

Although MTL models may outperform STL models. This paper addresses the problem of building a robust Arabic stance detection model through the single-task learning approach. It allows a focused analysis of the linguistic and contextual features specific to the Arabic language without the potential complexities and resource demands associated with multi-task learning frameworks. This approach ensures that the model is specifically tailored to recognize nuances and subtleties in Arabic discourse, leveraging a rich dataset curated for stance detection to optimize accuracy and performance.

Our paper offers the following contributions:

Proceedings of The Second Arabic Natural Language Processing Conference, pages 842–846
August 16, 2024 ©2024 Association for Computational Linguistics

Target	Train				Test				Total
	#Tweets	%Favor	%Against	%None	#Tweets	%Favor	%Against	%None	
COVID-19 Vaccine	1167	43.62	43.53	12.85	206	43.69	43.69	12.62	1373
Digital Transformation	1145	76.77	12.40	10.83	203	76.85	12.32	10.84	1348
Women Empowerment	1190	63.87	31.18	4.96	210	63.81	30.95	5.24	1400
All	3502	61.34	29.15	9.51	619	61.39	29.08	9.53	4121

Table 1: Distribution of the targets in the train and test datasets (as in Mawqif).

- We present an STL-based model that utilizes AraBERT v1 (Antoun et al., 2020) for detecting stances on the topics of the COVID-19 vaccine, digital transformation, and women's empowerment.

- We provide a detailed analysis and discussion of our model's performance across these topics.

The organization of this paper is as follows. In Section 2, we present prior and recent research on Arabic stance detection. Section 3 presents a comprehensive analysis of the dataset. Section 4 describes our proposed system and experimental setup. Section 5 presents and discusses our experimental results. In Section 6, we conclude and point out ideas for future search.

2 Related Works

Addressing the problem of Arabic stance detection using STL approach has drawn the attention of researchers recently. In (AlRowais and Al-Saeed, 2023), four stance detection models were built and compared using STL approach leveraging the transformers AraELECTRA (Antoun et al., 2021), MARBERT (Abdul-Mageed et al., 2021), AraBERT v1 (Antoun et al., 2020) and Qarib (Abdelali et al., 2021). In (Alqurashi, 2022), more than 1.8 million tweets were collected to gauge public opinions on distance education in KSA during the 2020 academic year. Various machine-learning and deep-learning models were trained and compared using STL approach. Haouari and Elsayed (2023) constructed and released AuSTR, a small-sized dataset to address the problem of detecting the stance of authorities towards rumors in tweets. AraBERT (Antoun et al., 2020) is trained using STL approach to detect the stances in their work. MTL approach has been used lately to address the problem of Arabic stance detection in (Alturayeif et al., 2023).

3 Data

This section provides a comprehensive description of the dataset released by the StanceEval organizers. The dataset comprises over 4000 Arabic tweets, encompassing diverse tweets representing different stances: in favor, against, or neutral.

The tweets are targeted towards one of three topics: "COVID-19 vaccine", "digital transformation", and "women empowerment". Each tweet is not only labeled for stance but also for sentiment and sarcasm. This comprehensive labeling allows for a more in-depth study of how different opinions interact and for evaluating a model trained on multiple opinion dimensions in a multi-task paradigm.

3.1 Data Size

The Mawqif corpus comprises more than 4000 tweets, which are divided into train and test sets. Figure 1 illustrates the distribution of the tweets in the train and test sets. Table 1 details the distribution of the targets within the train and test sets.

From Table 1, it is evident that the target distributions in the testing set mirror those in the training set, ensuring consistency across both datasets. Additionally, the targets are evenly distributed among the different categories. While the topic of COVID-19 Vaccine exhibits a balanced distribution between the stances 'Favor' and 'Against,' the topics of Digital Transformation and Women Empowerment show a bias towards the 'Favor' stance.

The balanced distribution in the COVID-19 Vaccine category allows for evaluating model performance in distinguishing between opposing views. The imbalanced nature of the Digital Transformation and Women Empowerment categories could be challenging for models to learn and accurately predict the 'Against' and 'None' stances. Due to their lower representation, the 'None' stance category, presents a challenge for models to identify neutral or unclear opinions, which are less frequent.

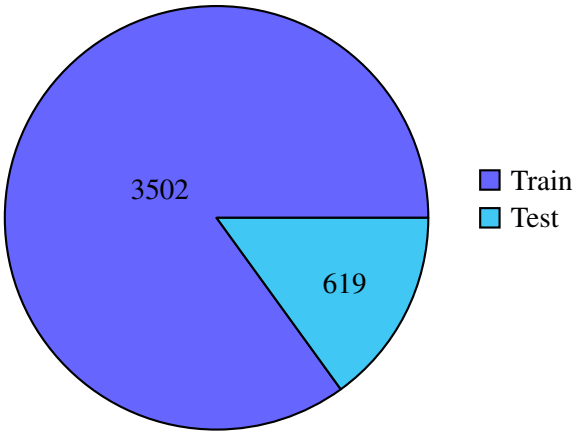

Figure 1: Statistics about tweets distribution in train and test sets

4 System Description

In this study, we aim to develop a stance detection model for Arabic text. To accomplish this, we have chosen to use a multi-class text classifier based on BERT models. This decision is based on the findings of (Alturayeif et al., 2022), who have demonstrated the superiority of BERT-based models in various text classification tasks.

ARABERT is an Arabic-pretrained language model based on Google's BERT architecture and uses the same BERT-Base configuration. The BERT architecture consists of multiple stacked Transformer encoders, each containing two sub-layers: a self-attention layer and a feed-forward layer. However, we fine-tuned the ARABERT v1 and v2 pre-trained models and built a standard pipeline under the Hugging Face framework. As per the exploration of evaluation metrics presented in Table 2, ARABERT v1 has already shown better results compared with ARABERT v2. So we decided to continue using the experiments with ARABERT v1, and as inspired by the empirical results. However, without detailed information about the exact differences between AraBERT V1 and V2, it's difficult for us to determine the exact reason. We consider this to be beyond the scope of this task.

4.1 Data Preprocessing

Our proposed system pipeline was implemented using Google Colab. The fine-tuning code for our model has been published and is available.[1]

The preprocessing of tweets involved preparing the Arabic texts by applying the following steps:

1. Removing numbers, all English letters, and newline.
2. Removing all words containing underscore, hashtag sign, and @ sign from the corpus.
3. removing diacritics, Tatweel, and non-Arabic characters.

Subsequently, the input text was tokenized using the BertTokenizer, which is a pre-trained tokenizer specifically designed for handling text data compatible with the BERT architecture. The tokenized text was then fed into the AraBERT model. For the classification task, the hidden representation corresponding to the special [CLS] token was extracted and employed as input to a feed-forward neural network layer coupled with a Softmax function.

4.2 Fine-Tuning Process

Given that pretrained models are trained on data from different domains and for different NLP tasks, it is necessary to retrain the model for the specific task using domain-specific data. This retraining process, known as fine-tuning, involves adjusting the model to the new task. Fine-tuning is a supervised learning technique where the weights from pretrained models are utilized as the initial weights for a new model tailored to the specific task (Imran and Amin, 2021). This method not only accelerates the training process but also results in state-of-the-art performance across various NLP tasks (Imran and Amin, 2021). It is an extremely effective strategy that leverages a pretrained model and adapts it to a task-relevant dataset.

In this project, retraining a pretrained model for a new task, such as those involving transformers, necessitates fine-tuning the training parameters. This fine-tuning process is essential to optimize the performance of each model for the specific task at hand.

4.3 Hyperparameters

During model training, hyperparameter tuning leverages computational infrastructure to evaluate multiple configurations of hyperparameters. This process helps identify the optimal hyperparameter values, thereby enhancing the prediction accuracy of the model. Hyperparameters play a crucial role in improving the performance of models in natural language processing (NLP).

In this task, we used TrainingArguments from the Hugging Face Transformers library, which in-

[1] https://github.com/ohamed/StanceEval24

Model	COVID-19 Vaccine	Digital Transformation	Women Empowerment
	F1-score		
CAMeLBERT-da	71.84	59.36	68.27
MARBERT	73.94	49.30	67.18
AraBERT	76.01	59.51	69.64
AraBERT-twitter	76.77	62.25	74.64
TAO AraBERT v2	73.05	60.14	82.02
TAO AraBERT v1	**90.36**	**92.36**	**91.34**

Table 2: Performance comparison of TAO stance detection based on the same test split against previous models reported in Alturayeif et al. (2022).

cludes several hyperparameters for configuring the training process. This configuration sets up primary training hyperparameters, including, among others:

- *Batch size*: The training batch size is set to 16, and the evaluation batch size is set to 32.
- *Epoch*: This refers to the number of times the entire dataset is passed through the model. We have set the number of epochs to 3.
- *Sequence length*: To retain as much data as possible, we have chosen the maximum sequence length available in the data, which is 128.
- *Learning Rate*: Initial learning rate for the optimizer. A good starting value for the learning rate in a multi-class classification task with transformer models is typically in the range of 2e-5 to 5e-5.

5 Results and Discussion

We divided the dataset into training, validation, and testing sets, allocating 70%, 20%, and 10% of the data, respectively. Our model, TAO, was then evaluated on the test set, achieving a macro-F1 score of 92.36 for digital transformation, 91.34 for women empowerment, and 90.36 for COVID-19 vaccine detection. Note that we performed our evaluation and calculated the macro F1-score using the Python script provided by the organizers. Thus, the macro F1-score is calculated as the average of the F1-scores for the "FAVOR" and "AGAINST" categories. The stance detection performance of TAO on our testing set, as compared with benchmark models reported by Alturayeif et al. (2022), is presented in Table 2. The results demonstrate that TAO outperforms the benchmark models CAMeLBERT-da, MARBERT, AraBERT, and AraBERT-twitter.

The performance of TAO declined when evaluated on the blind dataset provided by the shared task organizers. It achieved a macro-F1 score of 64.55 for digital transformation, 73.30 for women empowerment, and 70.51 for COVID-19 vaccine detection. Several factors may account for this performance decline. One major issue is the imbalanced distribution of 'Favor' and 'Against' instances within the Digital Transformation and Women Empowerment targets, as well as across the entire dataset. Additionally, the underrepresentation of the 'None' class posed a challenge for the model to learn and accurately predict these instances. Furthermore, the larger size of the blind test set compared to our own test set likely contributed to an increased prediction error.

6 Conclusion

In this paper, we present our BERT-based system for Arabic stance detection, developed for the ARABICNLP2024 STANCEEVAL shared task. Our model utilizes ARABERTV1, a pre-trained Arabic language model, within a single-task learning framework. The model achieves *macro-F1* scores of 73.30 for women empowerment, 70.51 for digital transformation, and 64.55 for COVID-19 vaccine detection.

For future work, we plan to explore the training of different systems, such as MARBERT, ARABERT-TWITTER, and XLM-R. Additionally, we aim to employ a multi-task learning (MTL) approach to further enhance the performance of our stance detection models.

Acknowledgments

The authors would like to thank Palestine Technical University—Kadoorie for providing support.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training BERT on arabic tweets: Practical considerations. *CoRR*, abs/2102.10684.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 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, Online. Association for Computational Linguistics.

Tahani Alqurashi. 2022. Stance analysis of distance education in the kingdom of saudi arabia during the COVID-19 pandemic using arabic twitter data. *Sensors*, 22(3):1006.

Reema Khaled AlRowais and Duaa AlSaeed. 2023. Arabic stance detection of covid-19 vaccination using transformer-based approaches: a comparison study. *Arab Gulf Journal of Scientific Research*.

Nora Saleh Alturayeif, Hamzah Luqman, and Moataz A. Ahmed. 2023. Enhancing stance detection through sequential weighted multi-task learning. *Soc. Netw. Anal. Min.*, 14(1):7.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label Arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184, 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. 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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Gilbert Badaro, Ramy Baly, Hazem M. Hajj, Wassim El-Hajj, Khaled Bashir Shaban, Nizar Habash, Ahmad Al Sallab, and Ali Hamdi. 2019. A survey of opinion mining in arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 18(3):27:1–27:52.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831, Online. Association for Computational Linguistics.

Imane Guellil, Houda Saadane, Faiçal Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *J. King Saud Univ. Comput. Inf. Sci.*, 33(5):497–507.

Fatima Haouari and Tamer Elsayed. 2023. Detecting stance of authorities towards rumors in arabic tweets: A preliminary study. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, volume 13981 of *Lecture Notes in Computer Science*, pages 430–438. Springer.

Abdullah Al Imran and Md Nur Amin. 2021. Deep bangla authorship attribution using transformer models. In *Computational Data and Social Networks - 10th International Conference, CSoNet 2021, Virtual Event, November 15-17, 2021, Proceedings*, volume 13116 of *Lecture Notes in Computer Science*, pages 118–128. Springer.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2024. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Trans. Knowl. Data Eng.*, 36(3):943–959.