Rasid at StanceEval: Fine-tuning MARBERT for Arabic Stance Detection

Nouf AlShenaifi
King Saud University
noalshenaifi@ksu.edu.sa

Nourah Alangari
King Saud University
nmalangari@ksu.edu.sa

Hadeel Alnegheimish
King Saud University
halnegheimish@ksu.edu.sa

Abstract

As social media usage continues to rise, the demand for systems to analyze opinions and sentiments expressed in textual data has become more critical. This paper presents our submission to the Stance Detection in Arabic Language Shared Task, in which we evaluated three models: the fine-tuned MARBERT Transformer, the fine-tuned AraBERT Transformer, and an Ensemble of Machine learning Classifiers. Our findings indicate that the MARBERT Transformer outperformed the other models in performance across all targets. In contrast, the Ensemble Classifier, which combines traditional machine learning techniques, demonstrated relatively lower effectiveness.

1 Introduction

In recent years, the rapid expansion of social media platforms, online news outlets, and digital communication has resulted in a significant rise in user-generated content. This unprecedented increase in online discourse necessitates automated tools and techniques to effectively analyze the opinions and attitudes expressed within these extensive streams of text. Stance detection, a crucial task in Natural Language Processing (NLP), seeks to identify a writer's position or perspective on a specific topic or entity by examining their written text and social media activity, such as preferences and connections (Küçük and Can, 2020). Stance detection differs from sentiment analysis in both focus and application. While sentiment analysis is concerned with identifying the emotional tone of a text, classifying it as positive, negative, or neutral, stance detection aims to determine the author's precise viewpoint or attitude toward a particular target or topic, identifying whether they are in favor, against, or neutral (Pang et al., 2008).

In this paper, we present our contributions to the shared task focused on developing models for detecting writers' stances (Favor, Against, or None) towards three selected topics: COVID-19 vaccine, digital transformation, and women empowerment (Alturayeif et al., 2024). We evaluate three models for stance detection in the context of Arabic micro-blogs, namely fine-tuning MARBERT (Abdul-Mageed et al., 2021), fine-tuning AraBERT Transformer (Antoun et al., 2020), and an ensemble classifier. The task is challenging due to the unique nature of the Arabic language, the informal writing styles prevalent on social media platforms, and the class imbalance in the dataset.

This paper is organized as follows: Section 2 provides a detailed description of the dataset used. Section 3 introduces the proposed models for stance detection in the Arabic language. Section 4 discusses the results obtained from our experiments. Finally, Section 5 presents the conclusion.

2 Data

This shared task is based on the MAWQIF (Alturayeif et al., 2022) dataset, an Arabic target-specific stance detection dataset collected from Twitter (now X). Given three specific targets, it contains multiple labels for each tweet, describing the polarities of stance, sentiment and sarcasm. Figure 1 shows word clouds for each of the targets.

Two data splits are provided for this task, Train and Blind Test. Train contains 3,502 samples for all targets, with 1167 for the Covid Vaccine target, 1145 for the Digital Transformation target and 1190 for the Women Empowerment target. Figure 2 shows the division of stance labels by target. We notice that the stance for Covid Vaccine is equally divided between in favor and against, whereas Digital Transformation has tweets mostly in favor. We further split Train with proportions 0.8 and 0.2 to generate a Validation set. We used the raw text for fine-tuning and testing our models, without any preprocessing, since normalizing and cleaning the

Proceedings of The Second Arabic Natural Language Processing Conference, pages 828–831

Figure 1: Word clouds for the text in the samples of each target in the training set.

Figure 2: Count of samples for each stance label by target in the training set.

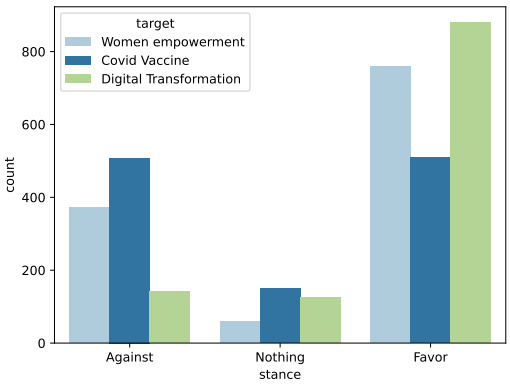

Table 1: Macro F1-score across all targets for various transformers and ML models

Model	Macro F1-score across all targets
Fine-tuning MARBERT	0.7566
Fine-tuning Arabert	0.7344
Ensemble Classifier	0.7033

text did not improve the the models' performance on this task.

3 System

Our submission involved fine-tuning a transformer-based BERT language model with a 3-way classification layer on top. We used a model that has been trained on both Dialectical Arabic (DA) and Modern Standard Arabic (MSA), MARBERT (Abdul-Mageed et al., 2021). We chose this model because its pretraining data included tweets, which contain informal and dialectical arabic, similar to the StanceEval task. The MARBERT model has the same architecture of $\text{BERT}_{\text{Base}}$, which contains approximately 163M parameters, namely, 12 layers, 768 hidden size, and 12 heads (Abdul-Mageed et al., 2021).

We used the huggingface library to fine-tune and evaluate our models (Wolf et al., 2020). We trained the model for 5 epochs with a 0.00001 learning rate, and 64 sample batch size. The maximum length for each sample was set to 256 tokens.

Moreover, we trained two additional models as baselines: a fine-tuned AraBERT (Antoun et al., 2020) and an ensemble classifier. The Ensemble Classifier combines Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes, augmented with a TF-IDF Vectorizer utilizing both word n-grams (n=1,2) and character n-grams (n=2,5).

4 Results and Discussion

In an effort to enhance stance detection in the Arabic language for participation in the Stance Detection in Arabic Language Shared Task, this work evaluated three distinct models: a fine-tuned MARBERT model (Abdul-Mageed et al., 2021), a fine-tuned Arabert model (Antoun et al., 2020), and an Ensemble Classifier. The evaluation of these models was conducted using the script provided by the organizers of the shared task, emphasizing precision, recall, and the macro F1-score across

Table 2: Performance results of various transformers and ML models

Model	Target	Precision	Recall	Macro f1-score
Fine-tuning MARBERT	**Women empowerment**	0.8037	0.8412	0.8191
	Covid Vaccine	0.7045	0.7389	0.7112
	Digital Transformation	0.7126	0.7687	0.7394
Fine-tuning Arabert	**Women empowerment**	0.7527	0.7368	0.7434
	Covid Vaccine	0.7242	0.6834	0.7030
	Digital Transformation	0.7452	0.7687	0.7568
Ensemble Classifier	**Women empowerment**	0.7509	0.7711	0.7601
	Covid Vaccine	0.69	0.7333	0.6987
	Digital Transformation	0.6397	0.6679	0.7033

various targets. According to Table 1, the MARBERT Transformer model, fine-tuned for this task, emerged as the top performer, achieving a macro F1-score of 0.7566 across all targets, including Women empowerment, Covid Vaccine, and Digital Transformation. This score highlights its efficacy in accurately determining stance. In comparison, the AraBERT Transformer, which was also fine-tuned, achieved a close macro F1-score of 0.7344. Meanwhile, despite its novel integration of various traditional machine learning methods, the Ensemble Classifier achieved a lower score of 0.7033.

As illustrated in Table 2, the MARBERT Transformer model for the Women Empowerment target demonstrated superior precision (0.8037) and recall (0.8412), achieving the highest F1-score of 0.8191 across all models and targets. The AraBERT and Ensemble models also delivered strong performances, with F1-scores of 0.7434 and 0.7601, respectively. In the case of the Covid Vaccine target, the MARBERT model outperformed others with an F1-score of 0.7112, demonstrating well-balanced precision and recall. Conversely, the AraBERT model demonstrated decreased precision, resulting in a reduced F1-score of 0.7030, whereas the Ensemble model, though exhibiting strong recall, fell short in precision, recording an F1-score of 0.6987. For the Digital Transformation target, the AraBERT model performed exceptionally well, achieving a high F1-score of 0.7568. The MARBERT Transformer closely followed with a macro F1-score of 0.7394, whereas the Ensemble model, facing more significant challenges, scored only 0.7033. The results reflect the difficulties encountered by traditional machine learning models in handling complex NLP tasks, in contrast to deep learning models that are advantaged by contextual embeddings.

5 Conclusion

In this paper, we presented our submitted models for the Stance Detection in Arabic Language Shared Task, where we evaluated three distinct models: a fine-tuned MARBERT model, a fine-tuned AraBERT model, and an Ensemble Classifier. Our evaluation revealed that the MARBERT Transformer consistently outperformed the others across various targets.

The AraBERT Transformer also showed competitive performance but fell slightly short of MARBERT's overall efficacy. The Ensemble Classifier, which integrated various traditional machine learning techniques, achieved lower performance. This result underscores the challenges traditional machine learning models face in handling the complexity of stance detection tasks compared to deep learning models with contextual embeddings.

Future research should focus on enhancing these models, integrating additional linguistic features, enlarging the dataset, and addressing the existing class imbalance to further improve stance detection accuracy in Arabic language.

References

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.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024: The first arabic stance detection shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

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 *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.