Team_Zero at StanceEval2024: Frozen PLMs for Arabic Stance Detection

Omar Galal
Computer Engineering Department
Cairo University
omargalal@eng.cu.edu.eg

Abdelrahman Kaseb
Computer Engineering Department
Cairo University
abdelrahman.kaseb@eng.cu.edu.eg

Abstract

This research explores the effectiveness of using pre-trained language models (PLMs) as feature extractors for Arabic stance detection on social media, focusing on topics like women empowerment, COVID-19 vaccination, and digital transformation. By leveraging sentence transformers to extract embeddings and incorporating aggregation architectures on top of BERT, we aim to achieve high performance without the computational expense of fine-tuning. Our approach demonstrates significant resource and time savings while maintaining competitive performance, scoring an F1-score of 78.62 on the test set. This study highlights the potential of PLMs in enhancing stance detection in Arabic social media analysis, offering a resource-efficient alternative to traditional fine-tuning methods.

1 Introduction

Social media has become a pivotal platform for public discourse, providing a dynamic space where individuals express opinions and engage in debates on a multitude of topics. Stance detection, a branch of natural language processing, plays a crucial role in analyzing these interactions by identifying the positions individuals hold towards specific issues. Stance detection is particularly significant in understanding public sentiment and the social dynamics around key topics such as women empowerment, COVID-19 vaccination, and digital transformation. Accurate stance detection enables researchers and policymakers to gauge the prevailing attitudes and respond appropriately, fostering informed decision-making and promoting constructive dialogue. In the context of Arabic social media, stance detection is essential for capturing the unique cultural and linguistic nuances that shape public opinion in the Arab world, thereby enhancing the relevance and impact of social media analytics.

In the era of Large Language Models (LLMs) such as LLama (et al, 2023) and Pre-trained Language Models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), the NLP community goes more towards using such language models to tackle several downstream tasks such as stance detection. PLMs such as BERT are usually pre-trained on a general language modeling task (e.g. Masked Language Modeling) to give the model a pretty good understanding of the language in general. To use the model then on a downstream task, the model weights are updated on a dataset annotated on that task. Such approaches of pre-training language models and then fine-tuning them on downstream tasks have proved to work exceptionally well and boosted the performance on most NLP common tasks. They also have even been successfully applied to less common tasks like depression detection from social media (Kaseb et al., 2022).

Although boosting NLP tasks' performance through fine-tuning, the fine-tuning operation is costly in terms of the needed resources and time. Fine-tuning PLMs usually requires powerful GPUs to train all the PLM weights on the downstream task and relatively high training time compared to simpler machine learning algorithms. In this context, some approaches proposed parameter-efficient fine-tuning approaches (Han et al., 2024) that mainly aim to minimize the number of parameters being updated on the downstream task. Other approaches (Galal et al., 2024b) (Galal et al., 2024a) showed that PLMs can be used as feature extractors only without the need to fine-tune them. This is done by adding an extra aggregation architecture on top of the PLM with a relatively small number of parameters compared to the PLM. These aggregation architectures help extract sentence embeddings that are more powerful for text classification tasks.

This work aims to explore the effectiveness of using PLMs as feature extractors - without any fine-tuning - on Arabic stance detection. The major goal of using PLMs as feature extractors is

Proceedings of The Second Arabic Natural Language Processing Conference, pages 783–787

Target	#Tweets	%Favor	%Against	%None
Women Empowerment	1190	63.87	31.18	4.96
COVID Vaccine	1167	43.62	43.53	12.85
Digital Transformation	1145	76.77	12.40	10.83
All	3502	61.34	29.15	9.51

Table 1: Mawqif training set statistics.

eliminating the need to fine-tune them and hence saving computational and time resources. Firstly, we explore the sentence transformers (Reimers and Gurevych, 2019) to extract sentence embeddings for the Arabic tweets. Secondly, we follow some of the approaches proposed by (Galal et al., 2024b) by adding aggregation architectures on top of BERT to get a more powerful sentence embedding for stance detection. Our work was submitted to the StanceEval (Alturayeif et al., 2024) competition and ranked fifth out of 16 teams. This shows how competitive PLMs can be used as feature extractors and gain a competitive performance besides saving the fine-tuning costs.

2 Data

To train our models on Arabic stance detection, we used the Mawqif dataset (Alturayeif et al., 2022). The dataset consists of 3502 Arabic tweets for training and 619 for testing spanning three topics: women empowerment, covid vaccine, and digital transformation. The dataset annotations for the stance are either "FAVOR", "AGAINST", or "NONE". The tweets are also annotated for sentiment and sarcasm to allow the capabilities of multi-task learning approaches. Table 1 shows the Mawqif training set's stance label distribution across the different targets.

3 Methodology

This section illustrates the approaches for stance detection. The approaches we followed can be classified into three main approaches: sentence transformers-based approaches, BERT-based embeddings, and parallel sum architecture on top of BERT.

3.1 Sentence Transformers

In this approach, the tweets are fed to a sentence transformer (Reimers and Gurevych, 2019) to extract sentence embeddings for these tweets. These embeddings are then used to train a logistic regression model (Cox, 1958) to predict the tweet stance. Logistic regression implementation was used from scikit-learn (Pedregosa et al., 2011) with a learning rate of 0.001 and a maximum number of iterations of 100. Since we are dealing with Arabic tweets, we used multilingual sentence transformers (Reimers and Gurevych, 2020). All model weights are available at Hugging Face (Wolf et al., 2019). Here are the models used in this research:

- paraphrase-multilingual-mpnet-base-v2
- distiluse-base-multilingual-cased-v1
- paraphrase-xlm-r-multilingual-v1
- LaBSE (Feng et al., 2020)
- distiluse-base-multilingual-cased-v2
- paraphrase-multilingual-MiniLM-L12
- use-cmlm-multilingual
- multi-qa-mpnet-base-dot-v1
- all-mpnet-base-v2

3.2 BERT-based Embeddings

In this part, BERT-based pre-trained language models were used as sentence embedding extractors. Following the original BERT paper, we used the [CLS] output embedding of BERT as the sentence embedding. Following the findings of (Galal et al., 2024b), we also averaged all BERT output embeddings to get a more representative sentence embedding keeping the PLM frozen. These embeddings are then used to train a logistic regression model to predict the stance of the tweet with the same settings defined in 3.1. Since there is a variety of pre-trained BERT models in Arabic, we worked with the most common Arabic BERT models. The following list sums up the models used in our work:

- AraBERTv0.2-Twitter (Antoun et al.)
- MARBERT (Abdul-Mageed et al., 2021)
- Qarib (Abdelali et al., 2021)

- ARBERT (Abdul-Mageed et al., 2021)
- Arabic-BERT (Safaya et al., 2020)

3.3 Parallel-Sum on Top of BERT

The Parallel-Sum (P-SUM) architecture was first proposed by (Karimi et al., 2020) to fine-tune BERT for aspect-based sentiment analysis. (Galal et al., 2024b) then showed that this architecture improves BERT's performance on Arabic sentiment analysis and sarcasm detection keeping BERT parameters frozen. They also showed an improvement in the architecture to work in the multi-task learning setting like (Kaseb and Farouk, 2022). Figure 1 shows the P-SUM architecture to be trained on tasks: stance detection and sentiment analysis. The base model used is MARBERT. The training was done for 5 epochs with a learning rate of $5e-4$. The best-performing checkpoint on the validation set was picked.

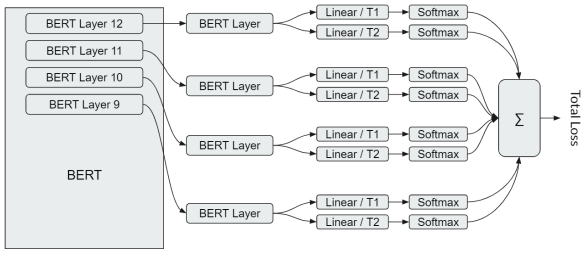

Figure 1: Multi-task learning with P-SUM (Galal et al., 2024b).

4 Experimental Results

4.1 Evaluation Metric

To evaluate our models, we used the same evaluation metric defined by the competition organizers which is the average of the F1-scores for the "FAVOR" and "AGAINST" categories. This metric is calculated for each target separately and then the final metric is computed across all targets by averaging the separate macro average F1-scores. The metric will be referred to as F_{avg2}

4.2 Results

Since the official test set was not released, the available dataset was randomly shuffled and 500 examples were picked to represent an internal test set to evaluate our models.

Table 2 shows the models' results on the internal test set. The table is split into three parts separated by a horizontal line. Each part represents a major approach. The first part represents the sentence transformers approach. The second part represents the BERT-based embeddings approach where each model is tried by either taking the [CLS] embedding or averaging all the output embeddings. The third one represents the P-SUM approach. Although the sentence transformers are pretrained to extract a sentence embedding, it can be noticed the performance gap between the sentence transformers approaches and BERT-based approaches. This performance gap is expected as all of the used sentence transformers are multilingual transformers but all BERT-based models are pre-trained on Arabic language specifically.

Table 2 also shows that there are some Arabic PLMs that perform better than others. This happens because of the data used to pre-train these models. For example, both MARBERT and AraBERT-Twitter are pre-trained on massive Arabic data containing a lot of tweets and hence perform better than the other models on the dataset we are working on.

Model	F_{avg2}
paraphrase-multilingual-mpnet-base-v2	**70.6**
distiluse-base-multilingual-cased-v1	68.3
paraphrase-xlm-r-multilingual-v1	65.5
LaBSE	64.1
distiluse-base-multilingual-cased-v2	63.6
paraphrase-multilingual-MiniLM-L12	63.4
use-cmlm-multilingual	62.3
multi-qa-mpnet-base-dot-v1	57.7
all-mpnet-base-v2	51.8
AraBERTv0.2-Twitter$_{CLS}$	74.9
AraBERTv0.2-Twitter$_{AVG}$	76.1
MARBERT$_{CLS}$	71.5
MARBERT$_{AVG}$	**76.7**
Qarib$_{CLS}$	68.8
Qarib$_{AVG}$	70.2
ARBERT$_{CLS}$	66.2
ARBERT$_{AVG}$	67.8
Arabic-BERT$_{CLS}$	63.3
Arabic-BERT$_{AVG}$	66.6
P-SUM-MTL	**76.3**

Table 2: Stance detection results on the internal test set.

4.3 Official Submission

For the competition's final submission on the blind official test set, we took the majority voting of the top three performing models: MARBERT$_{AVG}$, P-

SUM-MTL, and AraBERTv0.2-Twitter$_{AVG}$. Table 3 shows the detailed results per target and the final scoring overall targets on the test set. Our submission ranked fifth in the competition out of 16 teams which shows how effectively PLMs can be used as feature extractors only without the extra cost of fine-tuning while still performing competitively.

Target	F_{avg2}
Women Empowerment	85.99
Covid Vaccine	73.08
Digital Transformation	76.8
All	78.62

Table 3: The official results of our submission.

5 Discussion

A major challenge we faced while tackling the problem is the overfitting. The relatively small size of the training set is the major cause of this problem. It can also be seen that the dataset is already distributed across three targets which makes the size per each target very small. To overcome this issue, we followed several approaches such as early stopping.

Another challenge was the choice of either training on the whole dataset or training several models per target. We tried both paths for sentence transformers and found that training on the whole dataset is better. Our intuition for that is firstly training per target makes the training set much smaller for the model to learn from and hence increases the overfitting effect. The second intuition is that how people show their stance has a lot of similarities in the written tweet between the different targets and hence training on the whole dataset helps the model learn from all targets together.

For future work in this direction, we could increase the dataset size by making use of other published datasets for stance detection such as (Mubarak et al., 2022). Applying data augmentation to balance the dataset distributions across labels and to increase the overall dataset size is also believed to increase our models' performance. Additionally, employing active learning techniques to select more informative and diverse data samples can further enhance the model's accuracy and efficiency from the new dataset (Kaseb and Farouk, 2023).

6 Conclusion

This research underscores the vital role of social media in public discourse and the importance of stance detection in understanding public sentiment towards critical issues such as women empowerment, COVID-19 vaccination, and digital transformation. Our study demonstrated that pre-trained language models (PLMs) when used as feature extractors, can effectively enhance stance detection without the resource-intensive process of fine-tuning. By leveraging sentence transformers and incorporating aggregation architectures on top of BERT, we achieved competitive performance, evidenced by our strong showing in the StanceEval competition with an F1-score of 78.62. This approach not only saves significant computational resources and time but also opens new avenues for efficient and scalable NLP applications in social media analytics.

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, 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 The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

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, Minnesota. Association for Computational Linguistics.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Omar Galal, Ahmed H Abdel-Gawad, and Mona Farouk. 2024a. Federated freeze bert for text classification. *Journal of Big Data*, 11(1):28.

Omar Galal, Ahmed H Abdel-Gawad, and Mona Farouk. 2024b. Rethinking of bert sentence embedding for text classification.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020. Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*.

Abdelrahman Kaseb and Mona Farouk. 2022. SAIDS: A novel approach for sentiment analysis informed of dialect and sarcasm. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 22–30, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abdelrahman Kaseb and Mona Farouk. 2023. Active learning for arabic sentiment analysis. *Alexandria Engineering Journal*, 77:177–187.

Abdelrahman Kaseb, Omar Galal, and Dina Elreedy. 2022. Analysis on tweets towards covid-19 pandemic: An application of text-based depression detection. In *2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 131–136.

Hamdy Mubarak, Sabit Hassan, Shammur Absar Chowdhury, and Firoj Alam. 2022. Arcovidvac: Analyzing arabic tweets about covid-19 vaccination. *arXiv preprint arXiv:2201.06496*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.