PICT at StanceEval2024: Stance Detection in Arabic using Ensemble of Large Language Models

Ishaan Shukla*, **Ankit Vaidya***, **Geetanjali Kale**
Pune Institute of Computer Technology
{ishaanshukla10, ankitvaidya1905}@gmail.com,
gvkale@pict.edu

Abstract

This paper outlines our approach to the StanceEval 2024 - Arabic Stance Evaluation shared task. The goal of the task was to identify the stance, one out of three (Favor, Against or None) towards tweets based on three topics, namely - COVID-19 Vaccine, Digital Transformation and Women Empowerment. Our approach consists of fine-tuning BERT-based models efficiently for both, Single-Task Learning as well as Multi-Task Learning, the details of which are discussed. Finally, an ensemble was implemented on the best-performing models to maximize overall performance. We achieved a macro F1 score of 78.02% in this shared task. Our codebase is available publicly[1].

1 Introduction

In recent years, the exponential growth of social media platforms, online news outlets and digital communication has also led to a surge of user-generated content. Effectively analyzing the opinions and attitudes expressed within such content to understand the perspective of the user regarding the topic is called Stance Detection, which is a critical task in the field of Natural Language Processing (NLP). Previous works like (Küçük and Can, 2020) have focused on multiple definitions of the task. The most widely used one (topic-phrase stance) is the stance of a text - favor, against, neutral is detected concerning a topic like 'Women Empowerment'.

Arabic is the fifth most spoken language in the world with over 420 million speakers worldwide. Stance detection has been extensively studied for English and other languages, but the lack of research on this topic for Arabic language exposes the gap in research. Reasons for this are challenges due to unique scripting, morphology and dialects, which increase the complexity of texts to be processed by machine learning models. One other reason is the limited data available in Arabic for stance detection. There are extremely limited datasets available like (Baly et al., 2018a) which covers topics like the war in Syria and has very few data points and (Khouja, 2020) which deals with news titles.

The StanceEval Shared Task (Alturayeif et al., 2024) at the ArabicNLP conference aims to detect writers' stances towards three selected topics like COVID-19 Vaccine, Digital Transformation and Women Empowerment on the MAWQIF dataset (Alturayeif et al., 2022). Earlier approaches like (Khouja, 2020) use character-level LSTMs and a multilingual variant of BERT (mBERT) (Devlin et al., 2019). Other approaches include the usage of BERT-based models for zero shot stance detection (Allaway and McKeown, 2020), using gradient boosting classifiers (Baly et al., 2018b) or using end-to-end memory networks (Mohtarami et al., 2018).

We built upon the approach used by (Alhindi et al., 2021) using language specific BERT based models like CamelBERT (Inoue et al., 2021), MARBERT (Abdul-Mageed et al., 2021) and AraBERT (Antoun et al., 2021) and some of their variants. We ranked 7th in this shared task, with a net macro-F1 of 78.02%. Individual macro-F1 scores for categories of 'Women empowerment', 'Covid Vaccine' and 'Digital Transformation' were 75.52%, 79.85% and 78.69% respectively. The subsequent sections describe our approach used for the task. We also analyze the performance of our pipeline and present ablations after rigorous experiments.

2 Data

The dataset contains 3502 examples belonging to 3 distinct topics - COVID-19 Vaccine, Women Empowerment and Digital Transformation. The distribution of examples across these topics is evenly

*first author, equal contribution

[1] http://github.com/ishaan-shukla10/StanceEval2024

Proceedings of The Second Arabic Natural Language Processing Conference, pages 837–841
August 16, 2024 ©2024 Association for Computational Linguistics

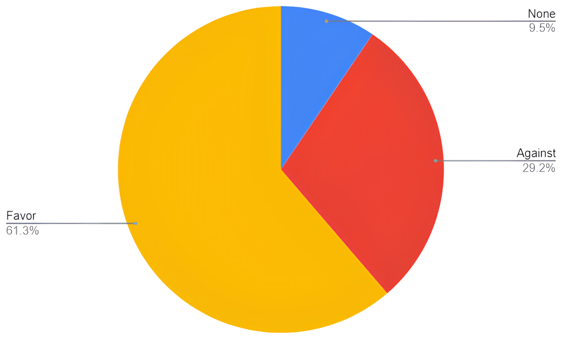

Figure 1: Class distribution of tweets

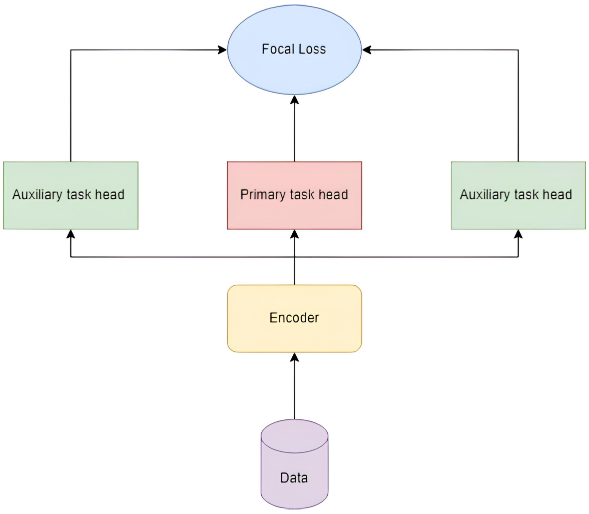

Figure 2: Multi-Task Learning Architecture

distributed. The possible values of stance are - Favor, Against and None. There is a slight class imbalance in this category. The class ’Favor’ has the highest number of examples (2148) whereas the class ’None’ has the least (333). The data distribution is illustrated in Figure 1. The dataset was a multi-label dataset with each data point having an annotation for stance, sarcasm and sentiment. We used a train-validation split of 80:20 on the given dataset. The training set had 2805 examples and the validation set had 697 examples. The test dataset contained 619 examples on which the system was finally evaluated. We observed a significant imbalance in the distribution of class labels. To mitigate that we used methods like weighted random sampling, however, there was no performance improvement even after using such methods. Hence, we did not use any data sampling methods while training the final system.

3 System Description

The problem of Stance Detection was a multi-class classification problem, in which the stance of given tweets was to be evaluated from a total of 3 classes. Both, Single-Task Learning (STL) and Multi-Task Learning (MTL) models were implemented in the system. We decided to use Multi-Task Learning because it can improve performance on the primary task while aiding in generalization due to the inductive bias introduced by the additional tasks (Caruana, 1997). To aid in MTL we used a focal loss in order to provide the right gradient signals to aid the primary task. The equation of the loss used is shown in Equation 1. The sum of the individual task-wise losses was divided by 6 for normalization. The primary task was stance detection and the auxiliary tasks were sarcasm detection and sentiment detection. It can be clearly observed from Table 2 and 3 that the MTL models have outperformed the STL models.

$$Loss = \frac{(sent + 4 * stance + sarc)}{6} \quad (1)$$

Then, depending on their performance on testing data, the best models were selected without any discrimination between STL and MTL models. We used models like MARBERT, MARBERTv2, AraBERT, CamelBERT, QARiB (Abdelali et al., 2021) their variants as the backbones in our system in the various experiments conducted. The pre-trained checkpoints for all the models were used through the HuggingFace (Wolf et al., 2020). We used basic pre-processing like the removal of links, mentions and English words. Also, the tweets contained emojis. To effectively deal with them we decided to replace each emoji with its description in Arabic. The models were then fine-tuned by connecting a linear layer to the pooler output was to the pooler output. Hyperparameter tuning was carried out to obtain the best possible model performance, by considering learning rate, weight decay, number of epochs, batch size and context length. For the top 3 models, the best set of hyperparameters we found are mentioned in Table 1. CrossEntropy loss was used for the individual tasks along with the Adam optimizer. The scoring metric of the task was macro F1 score for the ’Favor’ and ’Against’ stances only. The top models were selected based on the criterion of epoch-wise validation F1 score. We observed that an ensemble performed better than individual models. To account for all possible combinations from the top 5 MTL and STL models we tested all possible combinations of the models and accordingly chose the best system. Hard voting was used in the ensemble system. If there was no consensus in the system then the label predicted by the model with the highest validation F1-score was used as the final prediction.

Model name	**Batch Size**	**Learning Rate**	**Weight Decay**
AraBERTv2	20	10^{-5}	$5 * 10^{-6}$
CamelBERT	32	10^{-5}	10^{-6}
MARBERT	20	10^{-5}	10^{-6}

Table 1: Best set of hyperparameters for top 3 models

4 Results

A total of 13 models were used for experimentation. The models were the variants of the models mentioned above post-trained or fine-tuned on domain-specific data like tweets or dialect identification. The results of the top 5 models are depicted in Table 2 and Table 3 when trained as Single-Task Learning and Multi-Task learning models respectively. We observe clearly that models using the MTL paradigm show better performance as compared to STL models. However, in some cases, we observed that the MTL models overfit the training data as compared to their STL counterparts like CamelBERT. The bold entries from the tables show the models that were in the final system. The final ensemble system consisted of an MTL MARBERT, STL AraBERT and STL CamelBERT. The final results of our system can be seen in Table 4. Out of the 3 topics the system performed the best on the topic - COVID-19 Vaccine.

Model name	**Train F1**	**Val F1**
AraBERTv2	**98.38%**	**89.73%**
CamelBERT	**98.94%**	**89.23%**
MARBERTv2	97.55%	87.01%
QARiB_far	99.32%	86.44%
QARiB	99.87%	86.11%

Table 2: Results of STL models

Model name	**Train F1**	**Val F1**
MARBERT	**99.46%**	**89.65%**
AraBERTv2	99.79%	89.31%
MARBERTv2	99.43%	88%
CamelBERT	99.94%	85.29%
QARiB_far	98.61%	85.27%

Table 3: Results of MTL models

Team name	**WE F1**	**CV F1**	**DT F1**	**Overall F1-score**
PICT	75.52%	79.85%	78.69%	78.02%

Table 4: Results on final blind test set. WE - Women Empowerment, CV - COVID-19 Vaccine, DT - Digital Transformation.

5 Discussion

In this paper, we designed a system to perform stance detection on tweets belonging to various topics. To develop this system we performed rigorous experiments. The results of our system on the test dataset can be seen in Table 4. To further analyze our system we have plotted the confusion matrices of the system according to the various topics. One of the main points we observed is that the label ’None’ consistently gave the worst performance across all the 3 topics. We believe this is due to the fact there is a lack of any distinct words or semantic structure in the tweets belonging to this category unlike the categories of ’Favor’ and ’Against’. Another observation was that the models overfit when trained in an MTL setup as compared to the STL setup in some cases. We believe this can be mitigated by deciding the proper task order and giving the right importance to each task. The disparity in performance across topics despite of equal distribution of examples can be attributed to the fact that during pre-training the model encountered some topics more frequently and hence they have robust representations as compared to some other topics. One of the major drawbacks of our system is that the system tends to overfit on the data provided. To mitigate this we propose to use methods like finding the proper ordering of tasks so model can build robust representations and generalise better instead of overfitting and also trying to use more data. We can try to find additional data or try to use synthetic data. We can also try to improve ensembling approaches. One other drawback that needs to be addressed is that the methods we use require large amounts of data. In low-resource settings alternatives like synthetic data or efficient architectures need to be researched more so the task can be performed across multiple languages.

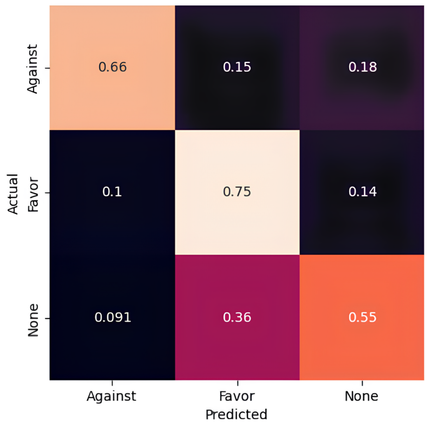

Figure 3: Confusion matrix of 3-way classification for ’Women empowerment’

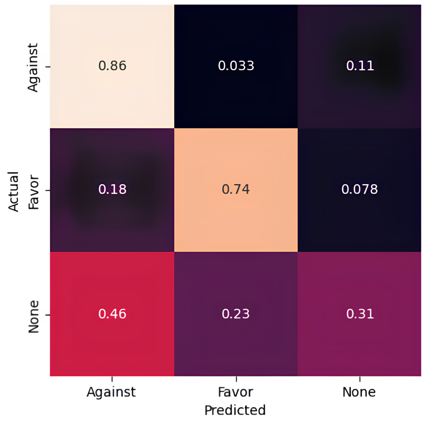

Figure 4: Confusion matrix of 3-way classification for ’Covid Vaccine’

6 Conclusion

This paper aims to describe our approach for the StanceEval 2024 Shared Task which involved classification of tweets towards 3 topics. We conducted experiments with multiple BERT-based models like AraBERT, MarBERT, CamelBERT and QARiB. We explored both single-task learning (STL) and multi-task learning (MTL) setups, finding that the MTL approach, which incorporated auxiliary tasks of sarcasm detection and sentiment analysis, outperformed the STL models. We also show that an ensemble of the best performing STL and MTL models achieved a strong macro F1-score of 78.02% on the evaluation dataset, with the models performing best on the COVID-19 Vaccine topic, followed by Digital Transformation and Women Empowerment. We also show that Multi-task Learning benefits the models and its drawbacks. We foresee several future directions for the work done. One direction can be to develop and use more sophisticated methods for ensembling and multi-task learning. Another direction is the development of systems that can achieve similar performance in low-resource settings either using synthetic data or by using efficient architectures.

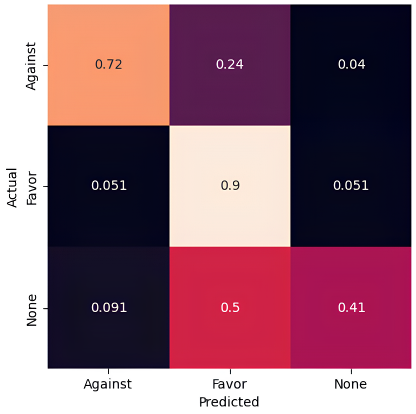

Figure 5: Confusion matrix of 3-way classification for ’Digital Transformation’

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.

Tariq Alhindi, Amal Alabdulkarim, Ali Alshehri, Muhammad Abdul-Mageed, and Preslav Nakov. 2021. AraStance: A multi-country and multi-domain dataset of Arabic stance detection for fact checking. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 57–65, Online. Association for Computational Linguistics.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, 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. 2021. Arabert: Transformer-based model for arabic language understanding.

Ramy Baly, Mitra Mohtarami, James Glass, Llu'is M'arquez, Alessandro Moschitti, and Preslav Nakov. 2018a. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '18, New Orleans, LA, USA.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28:41–75.

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.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, New Orleans, Louisiana. Association for Computational Linguistics.

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.