AlexUNLP-BH at StanceEval2024: Multiple Contrastive Losses Ensemble Strategy with Multi-Task Learning For Stance Detection in Arabic

Mohamed Badran, Mo'men Hamdy, Marwan Torki and **Nagwa El-Makky**
Computer and Systems Engineering Department
Alexandria University
{es-mohamedmostafabadran, es-moamenmohhamdy, mtorki, nagwamakky}@alexu.edu.eg

Abstract

Stance detection, an evolving task in natural language processing, involves understanding a writer's perspective on certain topics by analyzing his written text and interactions online, especially on social media platforms. In this paper, we outline our submission to the StanceEval 2024 task, leveraging the Mawqif dataset featured in The Second Arabic Natural Language Processing Conference. Our task is to detect writers' stances (Favor, Against, or None) towards three selected topics (COVID-19 vaccine, digital transformation, and women empowerment). We present our approach primarily relying on a contrastive loss ensemble strategy. Our proposed approach achieved an F1 score of 0.8438 and ranked first in the stanceEval 2024 task. The code and checkpoints are available at https://github.com/MBadran2000/Mawqif.git

1 Introduction

The increase in user-generated content is evident across news portals, user forums, blogs, and social media platforms such as Twitter, Facebook, and Instagram. This surge has led to the accumulation of vast amounts of textual data. These texts are readily available for analysis to serve diverse practical purposes (ALDayel and Magdy, 2021).

One such purpose is stance detection, also known as stance classification or stance prediction. For an input in the form of a piece of text and a target pair, stance detection is a classification problem where the stance of the author of the text is sought in the form of a category label from this set: Favor, Against, Neither (Mohammad et al., 2016).

Stance detection applications vary from trend and market analysis, obtaining user reviews for products, opinion surveys, targeted advertising, polling, predictions for elections and referendums, automatic media monitoring, and filtering out unconfirmed content for better user experience, to online public health surveillance (Küçük and Fazli, 2020).

Earlier research in stance detection primarily concentrated on analyzing debates within online forums. Recent efforts, however, have shifted towards analyzing social media platforms, notably Twitter. Unlike forum debates that are typically centralized, social media discussions on a given topic tend to be dispersed, although sometimes linked by hashtags. Despite this dispersion, the wealth of additional features inherent to social media renders it a valuable source of data for stance detection endeavors (ALDayel and Magdy, 2021).

While a significant portion of stance detection studies has focused on the English language, recent endeavors have aimed at developing stance detection datasets for other languages. However, no comparable initiatives appear to have addressed the challenges within the Arabic language (Alturayeif et al., 2023). Detecting stance in Arabic presents formidable challenges owing to its morphological intricacies and the diverse range of dialects present, which deviate from Modern Standard Arabic (Abdul-Mageed et al., 2018).

In this paper, we present our submission to Shared Task 4: StanceEval2024 (Alturayeif et al., 2024) in The Second Arabic Natural Language Processing Conference (ArabicNLP2024). The task focuses on stance detection in the Arabic language using the Mawqif dataset (Alturayeif et al., 2022). We tackled the problem by using multi-task model architecture, enhanced by multiple contrastive losses ensemble strategy. Our proposed approach achieves an overall F1 score of 0.8438, securing the first rank in the StanceEval 2024 task.

2 Data

For our experiments, we used Mawqif dataset only. It consists of 4,121 samples collected from Twitter platform. Each sample is annotated for three

Proceedings of The Second Arabic Natural Language Processing Conference, pages 823–827

	Women Empowerment	COVID Vaccine	Digital Transformation
Favor (%)	63.86	43.63	76.78
Against (%)	31.14	43.55	12.39
Neutral (%)	5	12.82	10.83

Table 1: Distribution of Stance Categories for each target

dimensions: stance, sentiment, and sarcasm. The considered targets in the dataset are COVID-19 vaccine, digital transformation, and women empowerment.

The dataset is distributed as follows across the 3 targets 33.3% (1373 tweets) for COVID-19 vaccine, 32.7% (1348 tweets) for digital transformation, and 33.97% (1400 tweets) for women empowerment The distribution of stance categories for each target is presented in Table 1.

We split the dataset into training, validation, and testing datasets with 70% for training, 15% for validation, and 15% for testing. To preprocess the data, we implemented several steps to ensure uniformity and enhance model performance. Firstly, we replaced Hindi numerals with their English counterparts to standardize numerical representations across the dataset. Additionally, we substituted mentions, URLs, and email addresses with special tokens to eliminate noise. Moreover, for models utilizing AraBERT V2, we removed emojis entirely as the model was pretrained without them.

3 System

In this section, we provide a detailed description of our system developed for stance detection in Arabic. The multitask model architecture including its associated losses is illustrated in Figure 1.

3.1 Baseline Models

We initiated our baseline by training a specific model for each target. For women empowerment and digital transformation, we used AraBERTv0.2-Twitter-base (Antoun et al.), which was pre-trained using the Masked Language Modeling (MLM) task on approximately 77 GB MSA Text plus approximately 60 million Arabic tweets. To address the COVID-19 vaccine target, we used AraBERT COVID-19 (Ameur and Aliane, 2021), a fine-tuned version of AraBERTv2. This model underwent fine-tuning using 1.5 million diverse Arabic tweets sourced from the "Large Arabic Twitter Dataset on COVID-19" (Alqurashi et al., 2020) to optimize its performance in addressing COVID-19-related tasks. Each target-specific model is trained on its target train data only.

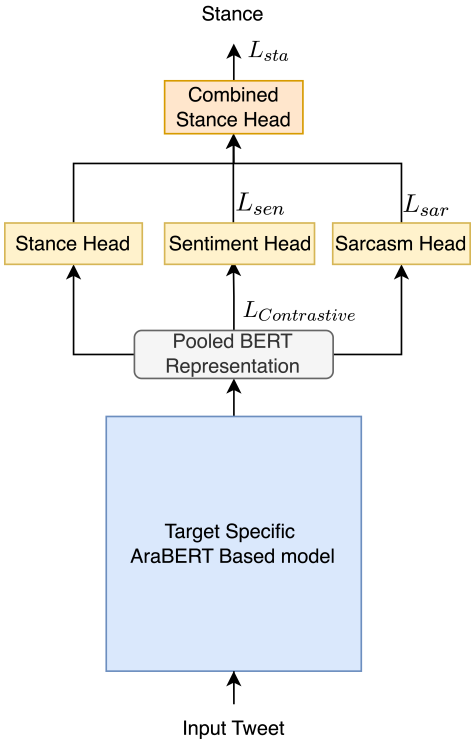

Figure 1: The multi-task model architecture with its associated losses.

3.2 Data Augmentation

To overcome the challenges posed by limited data and overfitting, we implemented data augmentation techniques during the data loading process at every iteration. Specifically, we employed two approaches: firstly, randomly removing 10% of the original comment words; and secondly, generating new tweets from existing ones by replacing 10% of the original tweet's words with their synonyms. These techniques aim to prevent the model from relying too heavily on specific patterns, thereby improving the generalization ability of the model (Ibrahim et al., 2018). Also, we used Dropout as an additional regularization technique.

3.3 Multi-Task Learning

Multi-task learning is training a single model to simultaneously perform multiple tasks, leveraging shared information to enhance overall performance (Caruana, 1997). In our approach, we employed multi-task learning to use the labeled data for sarcasm and sentiment analysis, aiming to improve stance detection (Alturayeif et al., 2023). To facilitate this, we incorporated three parallel heads,

Women Empowerment Macro F1 score	COVID Vaccine Macro F1 score	Digital Transformation Macro F1 score	Overall Macro F1 score across all targets
0.8855	0.8331	0.8127	0.8438

Table 2: Results on Official Test Set

Model	Modification	Women E. Macro F1 score	COVID V. Macro F1 score	Digital T. Macro F1 score	Overall Macro F1 score across all targets
1	Baseline	0.8424	0.7878	0.4345	0.6882
2	Model 1 with Data Augmentation and Dropout	0.8476	0.7893	0.6397	0.7589
3	Model 2 with Multi-Task Learning and Weighted Loss	0.8637	0.8343	0.7044	0.8008
4	Model 3 with Contrastive Loss Using L2 Distance	0.8803	0.7837	0.7876	0.8172
5	Model 3 with Triplet Margin Loss Using L2 Distance	0.8733	0.7813	0.7349	0.7965
6	Model 3 with FastAP Loss (Cakir et al., 2019)	0.8637	0.8184	0.7358	0.806
7	Model 3 with Tuplet Margin Loss (Yu and Tao, 2019)	0.8592	0.8043	0.7671	0.8102
8	Model 3 with Signal To Noise Ratio Contrastive Loss (Yuan et al., 2019)	0.855	0.7875	0.7318	0.7914
9	Ensemble of models 4+5+6+7+8	**0.8872**	**0.8279**	**0.8145**	**0.8432**

Table 3: Results on Our Test Set

each dedicated to one of the tasks, and subsequently combined their outputs into a unified classifier head for stance detection. We assigned a fixed weighting factor for each task to determine its relative importance, thereby directing the model's focus during training toward the most significant tasks. This multi-task framework enabled the model to benefit from the synergies between tasks, enhancing its ability in stance detection tasks. Additionally, to address data imbalances within each task, we employed weighted cross entropy loss, which helps mitigate the class imbalance problem, ensuring that the model learned from all instances effectively.

3.4 Contrastive Loss

We implemented contrastive loss (Khosla et al., 2020) as a key component of our training methodology. Contrastive Loss aims to minimize the distance between pooled BERT representations of similar sentence pairs while maximizing the distance between dissimilar pairs (Hussein et al., 2023). This approach is particularly beneficial for small datasets, as it helps mitigate imbalances and effectively increases the effective dataset size by an order of n^2, where n represents the number of samples (Shapiro et al., 2022). We applied contrastive loss specifically to stance labels only, as our primary task is stance detection. Our comprehensive loss function is illustrated in Equation 1

$$L = \lambda_1 L_{sta} + \lambda_2 L_{sen} + \lambda_3 L_{sar} + \alpha L_{Contrastive} \quad (1)$$

$Lsta$, $Lsen$, and $Lsar$ represent the weighted cross-entropy losses corresponding to the tasks of stance detection, sentiment analysis, and sarcasm detection respectively, Additionally, We incorporated contrastive loss function, training multiple models with different contrastive loss variants, and then combining them with ensemble. λ_1, λ_2, λ_3, and α are weighting factors that balance the importance of each loss function, The weighting factors are set to $\lambda 1 = 0.7$, $\lambda 2 = 0.2$, $\lambda 3 = 0.1$, and $\alpha = 1$.

3.5 Ensemble

We employed ensemble techniques to enhance the overall performance of our model. Ensemble learning is particularly effective when diverse models with different representations are combined (Dietterich, 2000). Our ensemble technique involves a straightforward approach where we combine the stance predictions from multiple models and select the stance that is most commonly predicted among them. This simple yet effective method allows us to leverage the strengths of each individual model to enhance overall system performance. To ensure that each model within our ensemble had distinct representations, we utilized different contrastive loss functions during training, which

encouraged each model to learn unique hidden representations for each class, thereby increasing the diversity among the ensemble members. This diversity among the models facilitated improved generalization and robustness, significantly enhancing the overall performance of our stance detection system.

4 Results

For our experimental settings, all experiments were conducted using a single RTX 3070 GPU. Each model was trained for 50 epochs with an early stopping patience of 5 epochs based on validation loss. The batch size was set to 32, and the maximum input sequence length was fixed at 128 tokens. Models incorporating dropout had a dropout probability of 0.1.

The macro F1 score is used to measure our results. This score is the average of the F1 scores for the "FAVOR" and "AGAINST" categories. This score is calculated for each target separately. Then, the overall macro F1 score is found by averaging these scores across all targets. Since each target in our approach has its own specific model, the F1 score is calculated individually for each model. Subsequently, the overall macro F1 score is computed by averaging these scores across all targets.

The results obtained from our comprehensive approach and modifications on the official test dataset are presented in Table 2. Additionally, Table 3 presents the performance on our test set, showcasing the effectiveness of various model enhancements. Model 1 in Table 3 serves as the baseline, trained without any modifications. Model 2 includes dropout and data augmentation, while Model 3 incorporates weighted loss and multi-task learning alongside the regularization techniques of Model 2. Models 4 to 8 further extend the modifications of Model 3 by training with different variants of contrastive loss functions each. Finally, Model 9 represents an ensemble of Models 4 to 8, demonstrating the collective performance of these augmented models.

5 Discussion

Our results on the official test set for the StanceEval task, as presented in Table 2. Specifically, we achieved an F1 score of 0.8855 for Women Empowerment, 0.8331 for COVID Vaccine, and 0.8127 for Digital Transformation. These individual target scores aggregated in an overall F1 score of 0.8438.

The results from our test set, as presented in Table 3, underscore the impact of each enhancement in our approach. Initially, our baseline model (Model 1) achieved an overall F1 score of 0.6882. However, with the introduction of dropout and data augmentation in Model 2, we effectively mitigated overfitting issues associated with the small dataset, resulting in a significant improvement to an overall F1 score of 0.7589.

In Model 3, multitask learning and weighted loss further improved our results, achieving an overall F1 score of 0.8008. The multitask approach enabled the model to simultaneously tackle related tasks, fostering more robust and generalized representations of the data. Addressing data imbalance, the weighted loss mechanism adjusted class contributions during training, prioritizing underrepresented classes. This combination of techniques proved highly effective in enhancing the model's performance.

Expanding our analysis to Models 4 to 8, where various contrastive loss functions were introduced, we observed mixed performance improvements. While some contrastive losses contributed positively to the overall performance, not all yielded significant gains. Nevertheless, the true efficacy of contrastive loss became evident in its impact on our ensemble strategy. By employing an ensemble of models trained with different contrastive losses, we achieved a notable overall F1 score of 0.8432. However, it's important to note that identifying the optimal combination of contrastive losses is a non-trivial task, requiring extensive experimentation and validation to determine the most effective ensemble configuration.

6 Conclusion

In this paper, we propose a multiple contrastive loss ensemble strategy for Arabic stance detection, which enabled us to achieve first place in the StanceEval 2024 task. Through our approach, we demonstrated that incorporating dropout and data augmentation mitigates overfitting while using a weighted loss effectively addresses data imbalance, and multi-task learning improves the generalization of our models. By integrating various contrastive losses, we enhanced the distinctiveness of hidden representations, contributing to a more effective ensemble. This work highlights the potential of ensemble techniques and contrastive learning to improve stance detection in Arabic texts.

References

Muhammad Abdul-Mageed, Haitham Alhuzali, and Mostafa Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. Enhancing stance detection through sequential weighted multi-task learning. *Social Network Analysis and Mining*, 14(1):7.

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.

Mohamed Seghir Hadj Ameur and Hassina Aliane. 2021. Aracovid19-mfh: Arabic covid-19 multi-label fake news and hate speech detection dataset. *Preprint*, arXiv:2105.03143.

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.

Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. 2019. Deep metric learning to rank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1861–1870.

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

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Mariam Hussein, Sarah Khaled, Marwan Torki, and Nagwa M El-Makky. 2023. Alex-u 2023 nlp at wojoodner shared task: Arabinder (bi-encoder for arabic named entity recognition). In *Proceedings of ArabicNLP 2023*, pages 797–802.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2018. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 875–878. IEEE.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Dilek Küçük and Cemal Aykut Nufus Fazli. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. Alexu-aic at arabic hate speech 2022: Contrast to classify. *arXiv preprint arXiv:2207.08557*.

Baosheng Yu and Dacheng Tao. 2019. Deep metric learning with tuplet margin loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6490–6499.

Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. 2019. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4815–4824.