ANLP RG at StanceEval2024: Comparative Evaluation of Stance, Sentiment and Sarcasm Detection

Amal Mezghani
ANLP Research group,
FSEG,
University of Sfax,
Tunisia
mezghani.amal@gmail.com

Rahma Boujelbane
ANLP Research group,
MIRACL Lab.,
University of Sfax,
Tunisia
rahma.boujelbane@fsegs.usf.tn

Mariem Ellouze
ANLP Research group,
MIRACL Lab.,
University of Sfax,
Tunisia
mariem.ellouze@fsegs.usf.tn

Abstract

As part of our study, we worked on three tasks: stance detection, sarcasm detection and sentiment analysis using fine-tuning techniques on BERT-based models. Fine-tuning parameters were carefully adjusted over multiple iterations to maximize model performance. The three tasks are essential in the field of natural language processing (NLP) and present unique challenges. Stance detection is a critical task aimed at identifying a writer's stances or viewpoints in relation to a topic. Sarcasm detection seeks to spot sarcastic expressions, while sentiment analysis determines the attitude expressed in a text. After numerous experiments, we identified Arabert-twitter as the model offering the best performance for all three tasks. In particular, it achieves a macro F-score of **78.08%** for stance detection, a macro F1-score of **59.51%** for sarcasm detection and a macro F1-score of **64.57%** for sentiment detection. .

Our source code is available at https://github.com/MezghaniAmal/Mawqif

1 Introduction

In a world where opinions abound and are constantly confronted, the ability to detect stances becomes crucial. Whether in the political arena, social debates or academic discussions, knowing how to identify stances becomes an essential tool for understanding underlying intentions and arguments. Detailed and nuanced understanding of natural language represents a major challenge for contemporary artificial intelligence, as highlighted by Chowdhary, 2020. It is important to recognize that decoding the subtleties of human language confronts us with the cultural and linguistic diversity that characterizes human communication. Researchers are thus stimulated to constantly push the limits of algorithmic understanding, by exploring the nuances of language in its various contexts. Every obstacle encountered, whether it is the variability of language, the ambiguity of expressions or the crucial need for context, represents an opportunity for learning and growth. Significant progress has been made in the field of stance detection. Recent work by Niu et al., 2024 introduced new datasets like MT-CSD and proposed innovative methods such as the Global-Local Attention Network (GLAN) , achieving a precision of only 50.47% in the stance detection task. Most stance detection research focuses on the English language, with relatively little research conducted on Arabic. Arabic is known for its syntactic complexity and rich vocabulary. Alhindi et al., 2021 focused on adapting models for better understanding and analyzing this language. They presented their new Arabic stance detection dataset (AraStance). They then evaluated this dataset, along with two other stance detection datasets, using several BERT-based models. Their best model achieved an accuracy of 85% and a macro F1 score of 78%. Recent work by Alrowais and AlSaeed, 2023 has focused on stance detection in Arabic tweets by comparing different stance detection models using four transformers: Araelectra, MARBERT, AraBERT, and Qarib. Their results show that the AraBERT model performed better than the other three models, with an F1 score of 70%. Jaziriyan et al., 2021 are working on target-based stance detection, which involves identifying a stance towards a specific target. Their approach resulted in a new corpus called ExaASC for the Arabic language. They used BERT to evaluate their corpus and achieved a macro F1 score of 70.69%.In addition to stance detection, understanding sarcasm represents a major challenge for sentiment analysis systems. Researchers like A. Galal et al., 2024 have worked on creating sarcastic corpora for Arabic and proposed hybrid deep learning approaches to improve sarcasm detection. Several works have also been carried out on the detection of sarcasm and sentiment, thus enriching our understanding of these complex aspects of

Proceedings of The Second Arabic Natural Language Processing Conference, pages 788–793

natural language. The shared task StanceEval Alturayeif et al., 2024 provides a platform to explore and perfect stance, sarcasm and sentiment detection techniques. It focuses on classifying stances in texts using different natural language processing (NLP) techniques to capture the subtlety of points of view expressed in various contexts. In this article, we look at the challenges and advancements in stance sensing. We explore the techniques, models, and approaches that have proven most effective in this area, while examining the specific challenges encountered when modeling these complex aspects of language. In addition to stance detection, we build models for sentiment and sarcasm detection tasks, which are crucial aspects of natural language analysis. We begin our article with a detailed description of the corpus of data used, highlighting the different classes represented and the distribution of labels. We also outline the data preprocessing pipeline we have in place to ensure the quality and relevance of our data before using it in our models. In **section 3** , we describe in detail the parameter tuning process we undertook to optimize the performance of our models. We explain the different steps of the process, including the pre-trained components used, as well as the hyperparameters adjusted to achieve the best possible performance. In **section 4**, we evaluate the performance of our models on the different components of our objectives. We present the evaluation metrics used and discuss the results obtained. In **section 5** we discuss the results obtained and propose avenues for improvement for the performance of our models in the tasks of position detection, sarcasm and sentiment analysis. Finally, we conclude our article by summarizing the main conclusions drawn from our study and proposing perspectives.

2 Data

In this work, we use MAWQIF dataSet Alturayeif et al., 2022 which is an important resource for the study of various topics, including "COVID-19 vaccination", "digital transformation" and "women's empowerment". It consists of 4121 entries across these three categories, with 1373 entries for COVID-19 vaccination, 1348 for digital transformation and 1400 for women's empowerment. This dataset is structured as a multi-label dataset, which means that an entry can be associated with multiple labels. Tags include "stance" which can be "Favor", "Against" or "None" "sentiment" which can be "Positive", "Negative" or "Neutral" as well as "sarcasm" which can be "Sarcastic " or "Non-sarcastic". These different dimensions allow us to carry out an in-depth analysis of the attitudes, opinions and linguistic nuances expressed in the texts. **Figure 1** shows the comments associated with their annotations in the MAWQIF dataset. Alturayeif et al., 2022.

Target	Tweet	Stance	Sentiment	Sarcasm
COVID-19 Vaccine	حاشا كورونا وطبنا منها ولله الحمد ومانحتاج تطعيم ولانحصنا أبدا We were diagnosed with Corona and recovered from it, thank God, we do not need a vaccination and we will never regret it	Against	Positive	No
Digital Transformation	مليون كتاب!! اين التحول الالكتروني للمناهج؟ كمية هدر سنوي للكتب مؤسفة تمنى احلال الاجهزة اللوحية بدلاً من الكتب Million books!! Where is the digital transformation of curricula? The amount of annual waste of books is unfortunate. We wish to replace books with tablets	Favor	Negative	No
Women Empowerment	😂😂#القبض_على_مدعي_النبوة فاهة تمكين المرأة غلط #Arrest_of_the_prosecutor_of_prophecy she misunderstod women's empowerment 😂😂	None	Neutral	Yes

Figure 1: Examples from MAWQIF dataset Alturayeif et al., 2022

Figure 2 shows the distribution of different labels in the training and testing parts of our dataset, highlighting an imbalance in the sarcasm label annotation. Moreover, this dataset comprises several Arabic dialects and subdialects, significantly complicating automated processing.

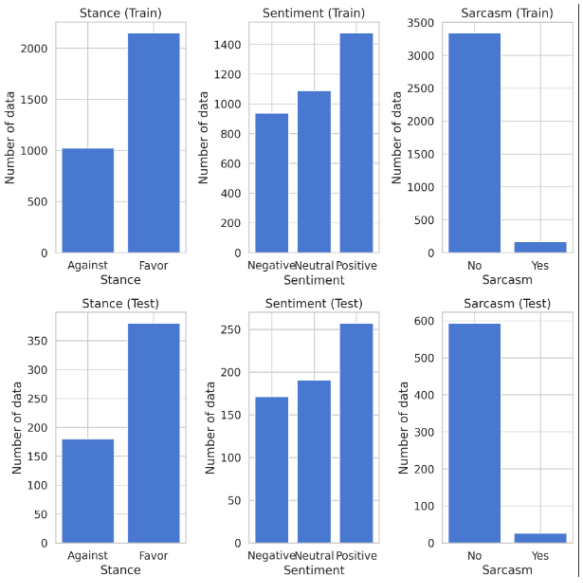

Figure 2: The distribution of different labels in the train and test sets

To analyze the distribution of texts based on their dialect, we employed the pre-trained model CAMeLBERT-MSA DID NADI Inoue et al., 2021. This model specializes in dialect identification, distinguishing between coarse-grained regional varieties and national varieties. It was developed by adapting the CAMeLBERT model originally designed for Modern Standard Arabic (MSA). **Figure 3** depicts the static distribution of dialects in our dataset, with Modern Standard Arabic (MSA) being the most dominant.

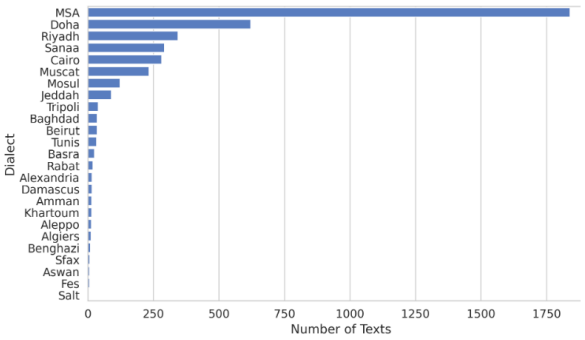

Figure 3: Distribution of dialects in the dataSet

Proposed Data Preprocessing Techniques

For the preprocessing phase, we used the Arabert-Preprocessor Antoun et al., 2021 class to perform various essential steps on the text. This includes removing HTML tags, replacing URLs, email addresses and attributions. We also used a pre-trained tokenizer to segment the text into meaningful tokens. This tokenizer, specially trained for the Arabic language, ensures adequate representation of the text by assigning a unique identifier to each token. Additionally, it offers advanced features such as adding special tokens, padding sequences, and creating attention masks to improve the quality of the input provided to the natural language processing model. Thus, the encoding process effectively converts the tokens into a numerical form suitable for use as input by the model. This preprocessing pipeline was applied to three different tasks: stance, sentiment, and sarcasm detection. However, for the latter task, since the dataset is unbalanced, we had to resort to data augmentation techniques. We used the Multi-dialect-Arabic-BERT model Talafha et al., 2020 to carry out two types of actions: "substitute" and "insert". As part of the "substitute" action, the augmentor replaces certain words in the input text with similar but different words. This makes it possible to diversify the content while retaining the general meaning of the text. For the "insert" action, the augmentor inserts additional words into the input text, which enriches the context and can potentially reinforce the sarcastic features of the text. In addition, we also applied a technique of removing certain words from the Arabic text as an augmentation method. These augmentation techniques were effective in increasing the number of data for the "Sarcastic" minority class in our datasets. This helped improve the model's performance in detecting sarcasm.

3 System

In this section, we fine-tuned four BERT-based models for each the three tasks: sarcasm detection, stance detection, and sentiment analysis.

The models used are as follows:

1. `CAMeLBERT-da`, a BERT-based model trained on 5.8 billion tokens from the Dialectal Arabic dataset;

2. `MARBERT`, a BERT-based model trained on 15.6 billion tokens from 1 billion Arabic tweets;

3. `AraBERT`, a model trained on 8.6 billion tokens from five datasets composed of Modern Standard Arabic (MSA) texts;

4. `AraBERT-twitter`, an extended AraBERT model (v0.2) trained on 60 million tweets in Arabic Alturayeif et al., 2022.

We fine-tuned each model for up to **20 epochs**, with a **maximum sequence length of 128**, a **batch size of 32**, an **AdamW optimizer with a learning rate of 2e-5**, and a **CosineAnnealingLR scheduler** to dynamically adjust the learning rate during model training. We initially divided the mawqif dataset into two parts: a training set and a validation set. This split was performed using the train_test_split function from scikit-learn, with stratification based on the "stance," "sarcasm," and "sentiment" columns to preserve the class distribution between the training and validation sets. The proportions are as follows: **82% of the data for the training set and 18% for the validation set**. This division ensures that model performance can be reliably assessed while maximizing the use of data for training. In addition to this split, we used a separate test set, loaded from a separate CSV file, which contains data specifically reserved for the

final evaluation of our model.The results reported in our study are based on the validation set for several reasons. Initially, we utilized the validation set to fine-tune hyperparameters and conduct cross-validation throughout the model training process. This approach enabled us to enhance the model's performance effectively while mitigating overfitting risks.Secondly, we continuously monitored and assessed the model's performance throughout its development, allowing for iterative adjustments and preliminary performance evaluations. Finally, the test set, sourced from a separate CSV file, was strictly reserved for the final evaluation after the model underwent full training and validation. This method ensures that the reported results are based on completely independent data, providing a more objective and reliable measure of the model's performance.

It's important to emphasize that this systematic approach was consistently applied to each task in our study—whether stance detection, sarcasm detection, or sentiment analysis. For every task, we followed the same protocol of data partitioning, hyperparameter optimization, and validation set evaluation before using the test set for the final assessment. This ensures a rigorous and standardized evaluation of our model's performance across various aspects of data analysis.

4 Results

The **AraBERT-twitter model** demonstrated the best overall performance as well as per-target performance. Tables **1**, **2** and **3** present the results of this model with the best configuration of parameters for stance, sarcasm and sentiment detection.Following an evaluation for stance detection, the overall F1 score reached **78.08%**. Thus, the Arabert-twitter model turns out to be more efficient in this task. Regarding sarcasm detection, despite the imbalance of classes for this task and an attempt to increase it was made, the model displays high accuracy, suggesting its ability to properly classify the majority (non-sarcastic) class. However, this result can be attributed to a possible class bias where the model is more inclined to predict the majority class, as well as to the distribution of predictions where a few correct predictions for the minority class can contribute to the overall accuracy , so it is essential to note that accuracy alone only for this task provides part of the picture and it is important to look at other metrics, such as the F1 score, which in this case reaches **59.51%**. An attempt to increase the corpus was undertaken, highlighting the major challenge of detecting sarcasm.Finally, for sentiment detection, the F1 score reached **64.57%**, demonstrating an intermediate performance of the model. These results highlight the need for further investigation, as in the area of sarcasm detection.

Table 1: **Performance Metrics for stance detection**

Metric	**Value**
Macro-average F1 over "For" and "Against" categories for **Women empowerment**	0.8207
Macro-average F1 over "For" and "Against" categories for **Covid Vaccine**	0.7892
Macro-average F1 over "For" and "Against" categories for **Digital Transformation**	0.7300
Macro F1-score across all targets	0.7808
Accuracy	0.7981

Table 2: **Performance Metrics for sarcasm detection**

Metric	**Value**
Macro-average F1 over "Sarcastic" and "Non-Sarcastic" categories for **Women empowerment**	0.5711
Macro-average F1 over "Sarcastic" and "Non-Sarcastic" categories for **Covid Vaccine**	0.7179
Macro-average F1 over "Sarcastic" and "Non-Sarcastic" categories for **Digital Transformation**	0.4963
Macro F1-score across all targets	0.5951
Accuracy	0.9483

Table 3: **Performance Metrics for sentiment detection**

Metric	Value
Macro-average F1 over "Neutral", "Positive", and "Negative" categories for **Women empowerment**	0.7504
Macro-average F1 over "Neutral", "Positive", and "Negative" categories for **Covid Vaccine**	0.6379
Macro-average F1 over "Neutral", "Positive", and "Negative" categories for **Digital Transformation**	0.5489
Macro F1-score across all targets	0.6457
Accuracy	0.6963

5 Discussion

Facing variable results despite using the same corpus of data for fine-tuning different BERT-based models, with superior performance in stance detection, we are confronted with a set of challenges and opportunities to enhance the effectiveness of our models.An imaginable strategy to improve performance is to explore combined or ensemble models. By merging individual models for each task, we could harness the strengths of each to achieve a more consistent overall performance. This approach could enable better utilization of the information available in the data and more robust classification on new examples. Furthermore, the adoption of multitask learning offers another intriguing avenue. By enabling the model to learn multiple tasks simultaneously, we could facilitate efficient knowledge sharing between them, potentially enhancing their respective performances. This would necessitate careful design of the model architecture and appropriate management of the weights assigned to each task to optimize results. By combining these approaches and tailoring them to our specific needs, we could hope to significantly enhance the performance of our models in stance detection, sarcasm detection and sentiment analysis tasks. This would require methodical work and rigorous evaluation, but the potential benefits in terms of precision and generalizability would certainly justify the effort.

6 Conclusion

In conclusion, our study highlights the crucial importance of understanding linguistic nuances in stance detection, sarcasm detection, and sentiment analysis through natural language processing (NLP) . By closely examining recent advancements in these domains, we underscored the remarkable progress made in processing conversational data, particularly in adapting models for Arabic language processing. Additionally, we shed light on the persistent challenges in sarcasm detection and sentiment analysis, while identifying innovative approaches aimed at enhancing model accuracy.

Through the use of fine-tuning techniques on BERT-based models, we have taken steps to maximize model performance in these essential tasks within the context of Arabic dialects. Our results emphasize the outstanding performance of the **Arabert-twitter model** in stance detection, achieving an impressive overall **F1 score of 78.08%**. It is crucial to highlight the pivotal role of StanceEval Alturayeif et al., 2024 in continuously improving stance, sarcasm and sentiment detection techniques, particularly in adapting these approaches to the specificities of Arabic dialects. These advancements represent a significant step towards a deeper and more precise understanding of human language, particularly in the realm of NLP applied to Arabic dialects.

References

Mohamed A. Galal, Ahmed Hassan Yousef, Hala H. Zayed, and Walaa Medhat. 2024. Arabic sarcasm detection: An enhanced fine-tuned language model approach. *Ain Shams Engineering Journal*, 15(6):102736.

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. *Preprint*, arXiv:2104.13559.

Reema 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 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. *Preprint*, arXiv:2003.00104.

K. R. Chowdhary. 2020. *Natural Language Processing*, pages 603–649. Springer India, New Delhi.

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. *Preprint*, arXiv:2103.06678.

Mohammad Mehdi Jaziriyan, Ahmad Akbari, and Hamed Karbasi. 2021. Exaasc: A general target-based stance detection corpus in arabic language. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 424–429.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. A challenge dataset and effective models for conversational stance detection. *Preprint*, arXiv:2403.11145.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T. Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. *Preprint*, arXiv:2007.05612.