dzStance at StanceEval2024: Arabic Stance Detection based on Sentence Transformers

Mohamed Lichouri
LCPTS, FGE-USTHB
Algiers-ALGERIA
mlichouri@usthb.dz

Khaled Lounnas
CRSTDLA
Algiers-ALGERIA
k.lounnas@crstdla.dz

Khelil Rafik Ouaras, Mohamed Abi and Anis Guechtouli
Algiers 01 University
Algiers-ALGERIA
{kh.ouaras, m.abi, a.guechtouli}@univ-alger.dz

Abstract

This study compares Term Frequency-Inverse Document Frequency (TF-IDF) features with Sentence Transformers for detecting writers' stances—favorable, opposing, or neutral—towards three significant topics: COVID-19 vaccine, digital transformation, and women empowerment. Through empirical evaluation, we demonstrate that Sentence Transformers outperform TF-IDF features across various experimental setups. Our team, dzStance, participated in a stance detection competition, achieving the 13th position (74.91%) among 15 teams in Women Empowerment, 10th (73.43%) in COVID Vaccine, and 12th (66.97%) in Digital Transformation. Overall, our team's performance ranked 13th (71.77%) among all participants. Notably, our approach achieved promising F1-scores, highlighting its effectiveness in identifying writers' stances on diverse topics. These results underscore the potential of Sentence Transformers to enhance stance detection models for addressing critical societal issues.

1 Introduction

Stance detection is a pivotal task in Natural Language Processing (NLP) that involves determining the position or attitude expressed in a text regarding a specific topic or entity. This task is crucial for a variety of applications, including sentiment analysis, opinion mining, and social media monitoring, where understanding public sentiment and opinion is essential (Alturayeif et al., 2022). With the exponential growth of user-generated content on social media and online news platforms, there is a pressing need to develop sophisticated tools that can analyze and interpret the myriad of perspectives present in these texts.

The Mawqif 2022 shared task, an initiative focused on Arabic stance detection, addresses this need by challenging participants to detect stances towards three contemporary topics: COVID-19 vaccine, digital transformation, and women empowerment. This task is significant not only because it addresses critical societal issues but also because it highlights the complexities of processing Arabic text, which is characterized by its rich morphology, diverse dialects, and intricate syntax (Alturayeif et al., 2022).

Traditional methods for stance detection have largely relied on feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF). These methods transform textual data into a numerical format, allowing machine learning models to process and analyze the text. Our initial experiments leverage TF-IDF features due to their simplicity and proven effectiveness in various text classification tasks (Abbas et al., 2019; Lichouri and Abbas, 2020).

Despite their utility, TF-IDF-based methods have limitations, particularly in capturing the deeper semantic relationships and contextual nuances within text. To address these limitations, recent research has increasingly focused on deep learning techniques. Long Short-Term Memory (LSTM) networks have demonstrated their ability to handle sequential data and capture dependencies in text, which are crucial for understanding stances (Alturayeif et al., 2023; Lai et al., 2020).

The advent of transformer-based models, such as BERT and Sentence Transformers, has revolutionized the field of NLP by offering robust methods to capture semantic and contextual information in text. These models utilize self-attention mechanisms to understand the relationships within text, making them particularly effective for nuanced tasks like stance detection (Reimers and Gurevych, 2019; Al-Dayel and Magdy, 2021). Transformers have set new benchmarks in various NLP tasks by leveraging their ability to generate dense, context-aware

representations of text, facilitating a deeper understanding of the underlying meaning and intent.

Several studies have explored the use of BERT and other transformers for stance detection, demonstrating their superiority over traditional methods in capturing subtle and complex stances in text. For example, Alshahrani et al. (AlDayel and Magdy, 2021) showed that BERT-based models outperform traditional approaches in detecting stances in English social media texts. However, these models often require extensive computational resources and fine-tuning, posing challenges for their application in resource-limited settings (Alturayeif et al., 2023).

Building upon these advancements, our study explores a comparative analysis between traditional feature extraction methods and modern deep learning approaches for Arabic stance detection. In our experiments, we employ both TF-IDF and Sentence Transformers to detect stances towards the selected topics. Our participation in the Mawqif 2022 shared task allowed us to evaluate these methodologies rigorously, where our team, dzStance, achieved competitive results across the three topics, demonstrating the effectiveness of our approaches.

In the following sections, we will delve deeper into our methodology and findings. Section 2 provides an overview of the Mawqif dataset used in our study. Section 3 describes our approach to stance detection, including the feature extraction techniques and model architectures we employed. Section 4 presents our experimental results and discusses their implications. Finally, Section 5 concludes the paper by summarizing the key insights and contributions of our work.

2 Dataset Description

The **Mawqif dataset** (Alturayeif et al., 2022), utilized in the **StanceEval 2024 shared task** (Alturayeif et al., 2024), serves as a crucial resource for advancing natural language processing (NLP) in the domain of stance detection. This dataset comprises over 4,000 annotated text samples that encapsulate diverse stances—*favorable*, *opposing*, or *neutral*—on pertinent topics such as COVID-19 vaccine, digital transformation, and women empowerment.

The significance of the Mawqif dataset lies in its ability to provide a comprehensive view of how different opinions and attitudes are expressed in Ara-

bic text. This makes it invaluable for researchers who aim to evaluate and enhance stance detection models. By leveraging such a dataset, one can explore and refine models to better understand and process nuanced stances within varied contexts.

Table 1 offers a detailed breakdown of the dataset, illustrating the distribution of tweets across the specified topics. The dataset comprises a total of 3,502 tweets, with the distribution as follows:

A closer look at the data statistics reveals notable class imbalances. For instance, the COVID-19 vaccine category includes nearly equal proportions of favorable (43.53%) and opposing (43.48%) tweets, with a smaller fraction being neutral (12.85%). In contrast, the digital transformation topic shows a predominance of favorable stances (76.77%), with fewer opposing (12.41%) and neutral (1.92%) tweets. Similarly, the women empowerment category also leans heavily towards favorable stances (63.95%), followed by opposing (31.18%) and neutral (4.96%) tweets.

Such imbalances can pose significant challenges for model training and evaluation. Models trained on datasets with skewed class distributions may become biased towards the majority classes, leading to suboptimal performance on minority classes. Therefore, addressing these imbalances is critical to ensure the development of robust and fair models. Techniques like data resampling, class weighting, and the use of advanced algorithms capable of handling imbalance are essential strategies to mitigate these effects and enhance overall model performance.

By understanding and leveraging the characteristics of the Mawqif dataset, researchers can effectively tackle the complexities of stance detection, contributing to the broader field of Arabic NLP and enabling more accurate and nuanced analysis of opinions and attitudes expressed in text.

3 Proposed System

In our proposed system [1], we explore two distinct methodologies for feature extraction: a weighted union of TF-IDF features (Lichouri et al., 2023) and Sentence Transformers. These techniques offer complementary advantages, leveraging the strengths of traditional feature representation and cutting-edge deep learning architectures.

Experiment 1 : Traditional Machine Learning with TF-IDF Features In our first experi-

---

[1]https://github.com/licvol/dzStanceEval_2024

Table 1: Data statistics of MAWQIF dataset.

Target	#Tweets	#Favor	#Against	#None
COVID-19 Vaccine	1167	508	507	152
Digital Transformation	1145	879	142	22
Women Empowerment	1190	761	371	59
All	3502	2154	1020	332

Id	Model	Text Feat	Configuration	Other	F1-score
1	LSVC	ngram_range=(1,4),(1,4),(1,4)	C=4	tw1	60.57
2	LSVC	ngram_range=(1,1),(1,1),(1,1)	C=4	tw1	61.15
3	LSVC	ngram_range=(1,5),(1,5),(1,5)	C=4	tw1	61.27
4	LSVC	ngram_range=(1,7),(1,7),(1,7)	C=4	tw1	61.32
5	LSVC	ngram_range=(1,10),(1,10),(1,10)	C=4	tw1	61.9
6	LSVC	ngram_range=(1,2),(1,2),(1,2)	C=4	tw1	62.1
7	LSVC	ngram_range=(1,3),(1,3),(1,3)	C=4	tw1	63.82
8	LSVC	ngram_range=(1,8),(1,8),(1,8)	C=4	tw1	63.92
9	LSVC	ngram_range=(1,9),(1,9),(1,9)	C=4	tw1	64.66
10	LSVC	ngram_range=(1,6),(1,6),(1,6)	C=4	tw1	66.2
11	LR	XLM-RoBERTa	max_iter=1000 multi_class='multinomial' solver='lbfgs'	na, re	**68.48**

Table 2: Obtained F1-score in the development phase. tw1:={0.85,0.85,0.65}, na:=normalize_arabic and re:=replace_emojis

ment, we focused on extracting features using the Term Frequency-Inverse Document Frequency (TF-IDF) approach, which is widely used in text classification tasks. We utilized scikit-learn's `FeatureUnion` module to combine different TF-IDF features, capturing both character-level and word-level information (Lichouri et al., 2021).

- **N-gram Range:** We experimented with various n-gram ranges to understand their impact on model performance. The ranges included:
  - (1,1): Unigrams
  - (1,2): Unigrams and bigrams
  - (1,3): Up to trigrams
  - (1,4): Up to 4-grams
  - (1,5): Up to 5-grams
  - (1,6): Up to 6-grams
  - (1,7): Up to 7-grams
  - (1,8): Up to 8-grams
  - (1,9): Up to 9-grams
  - (1,10): Up to 10-grams
- **Weighting:** To further enhance the TF-IDF features, we incorporated weighting schemes. The weights were varied

from 0.1 to 1.0 in steps of 0.1 to determine the optimal balance for capturing the intricacies of Arabic text. The best-performing weight, denoted as `tw1`, was identified and applied consistently across subsequent experiments (see Table 2).

- **Classifier:** For classification, we employed the Linear Support Vector Classifier (LSVC) with a regularization parameter set to $C = 4$. This choice was based on its ability to handle high-dimensional feature spaces effectively, which is crucial when dealing with the extensive n-gram features produced by TF-IDF.

This combination of weighted TF-IDF features and LSVC forms the baseline for our stance detection system, aiming to capture both the surface-level and deeper linguistic patterns in Arabic text.

Experiment 2 : Leveraging Pre-trained Language Models (PLMs) In the second experiment, we explored the use of advanced pre-trained language models (PLMs) to enhance stance detection capabilities further. These models are

pre-trained on vast amounts of text data and are adept at generating rich semantic representations of words and sentences.

- **Sentence Embeddings:** We utilized Sentence Transformers, specifically the **xlm-r-bert-base-nli-stsb-mean-tokens** model, which excels at producing dense vector embeddings that encapsulate the overall meaning of sentences. These embeddings were used as input features for the classification task.
- **Classifier:** For classification, we opted for a Logistic Regression (LR) model configured with the following hyperparameters:
  - `max_iter=1000`: This parameter sets the maximum number of iterations for the solver to converge.
  - `multi_class='multinomial'`: This setting enables the classifier to handle multiple classes simultaneously, which is crucial for stance detection where multiple stance labels exist.
  - `solver='lbfgs'`: The solver used for optimization, chosen for its efficiency in handling multiclass logistic regression problems.

This approach leverages the powerful representations learned by Sentence Transformers, which are adept at capturing semantic nuances and contextual relationships within the text. By integrating these embeddings with a logistic regression classifier, we aim to improve the model's ability to discern subtle stance indicators in Arabic text.

Combining these methodologies, our system aims to balance the strengths of traditional feature extraction with the advanced capabilities of modern pre-trained models. This hybrid approach is designed to address the linguistic complexity and variability of Arabic, providing robust stance detection across different contexts.

4   Results and Discussion

This section evaluates the performance of our stance detection system on the test set using the F1-score as the primary metric. We conducted two sets of experiments to compare the effectiveness of different feature extraction techniques and model configurations. Following these, we analyze our competitive performance in the stance detection challenge.

4.1   Baseline Experiment

To establish a benchmark, we employed a simple approach using Term Frequency-Inverse Document Frequency (TF-IDF) representation with unigram (1-gram) features and a Linear Support Vector Classifier (LSVC) with a linear kernel. This baseline model achieved an F1-score of 64.34%, providing a reference point for comparing the performance of more advanced models and feature combinations.

4.2   Experiment 1: Weighted Union of TF-IDF Features

In this experiment, we explored the impact of various n-gram ranges on the performance of the LSVC model. We employed the FeatureUnion module to create a weighted combination of TF-IDF features with different n-gram lengths. The weights were varied systematically from 0.1 to 1.0 to optimize feature importance. The optimal weight configuration (0.85, 0.85, 0.65) from this tuning was then used across all n-gram experiments.

We tested a range of n-grams from single-word (1-gram) up to ten-word sequences, examining the effects of character-level and word-boundary-aware features. The results show a consistent trend: the F1-score improves as the n-gram range increases, reaching a peak of 66.20% with six-grams (ngram_range=(1,6)). This suggests that incorporating up to six-word sequences captures essential context and relationships, enhancing the model's performance in stance detection tasks.

Interestingly, the performance slightly declines for n-grams longer than six-grams (e.g., ngram_range=(1,7) or ngram_range=(1,8)), possibly due to the introduction of noise or redundant information. These findings indicate that while expanding n-gram ranges can enrich the feature set, overly long sequences may adversely affect model accuracy.

4.3   Experiment 2: Sentence Transformers

The second experiment employed pre-trained language models to generate rich sentence embeddings. We used Sentence Transformers, specifically the 'xlm-r-bert-base-nli-stsb-mean-tokens' model, to create embeddings that encapsulate the

semantic meaning of each sentence. These embeddings were then fed into a Logistic Regression (LR) classifier for stance detection.

Configured with default hyperparameters—max_iter=1000, multi_class='multinomial', and solver='lbfgs'—and additional text preprocessing steps such as normalization and emoji replacement, the LR model achieved an F1-score of 68.48%. This score surpasses the highest result obtained in the TF-IDF-based LSVC experiments, highlighting the effectiveness of Sentence Transformers in capturing semantic relationships within text.

4.4 Competitive Performance Analysis

Our team, dzStance, participated in the StanceEval shared task, where the competition focused on detecting stances in various topical domains. Our performance across different topics was as follows: - **Women Empowerment**: We achieved the 13th position with an F1-score of 74.91% among 15 participating teams. - **COVID-19 Vaccine**: We ranked 10th, securing a 73.43% F1-score. - **Digital Transformation**: We placed 12th with a 66.97% F1-score. - **Overall Performance**: Combining all categories, dzStance ranked 13th overall with an F1-score of 71.77%.

These results reflect our system's ability to handle complex stance detection tasks, particularly in the context of the nuanced and diverse opinions expressed in the dataset. Despite the competitive nature of the task, our approach demonstrated robustness across different domains, indicating its potential for broader applications in stance detection.

4.5 Analysis and Insights

Our experiments underscore the importance of feature engineering and model selection in stance detection tasks. For the LSVC model, carefully selecting and combining n-grams up to six words proved most effective. This approach aligns with the need to capture both local and contextual information in Arabic text, characterized by its rich morphology and varying dialects.

On the other hand, the use of Sentence Transformers and Logistic Regression provided a significant performance boost. This suggests that leveraging pre-trained embeddings, which encode comprehensive semantic information, can substantially enhance the ability of stance detection models to interpret complex texts.

Overall, our competitive performance highlights areas for improvement but also demonstrates the potential of our methodologies. Future work could explore further fine-tuning of pre-trained models or combining TF-IDF and embedding-based approaches to harness the strengths of both methods, potentially leading to even greater improvements in stance detection performance.

5 Conclusion

In this paper, we introduced dzStance, our solution to the StanceEval 2024 shared task on Arabic stance detection. Leveraging Sentence Transformers in conjunction with Logistic Regression, our approach achieved competitive results with an overall average F1-score of 71.77%. This performance positioned us 13th among all participating teams, highlighting the effectiveness of advanced embedding techniques and robust classification algorithms in handling the complexities of Arabic stance detection. The success of our approach underscores the significance of pre-trained models like Sentence Transformers for capturing nuanced semantic relationships within Arabic text across diverse topics.

Looking ahead, further investigation into why the Logistic Regression model outperformed traditional methods such as Linear Support Vector Classification (LSVC) could yield insights through deeper hyperparameter tuning and broader evaluation on varied datasets. Additionally, exploring hybrid approaches that integrate TF-IDF features with advanced embedding models may offer enhanced model robustness and accuracy in Arabic natural language processing tasks. By making our code and methodologies openly accessible, we aim to foster reproducibility and encourage ongoing advancements in Arabic stance detection research, paving the way for more sophisticated and effective models in the future.

References

Mourad Abbas, Mohamed Lichouri, and Abed Alhakim Freihat. 2019. St madar 2019 shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 269–273.

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

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

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.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Mohamed Lichouri and Mourad Abbas. 2020. Speech-trans@ smm4h'20: Impact of preprocessing and n-grams on automatic classification of tweets that mention medications. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 118–120.

Mohamed Lichouri, Mourad Abbas, Khaled Lounnas, Besma Benaziz, and Aicha Zitouni. 2021. Arabic dialect identification based on a weighted concatenation of TF-IDF features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 282–286, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache, and Rachida Djeradi. 2023. Usthb at nadi 2023 shared task: Exploring preprocessing and feature engineering strategies for arabic dialect identification. *arXiv preprint arXiv:2312.10536*.

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.