StanceEval 2024: The First Arabic Stance Detection Shared Task

Nora Alturayeif[1], Hamzah Luqman[2,3], Zaid Alyafeai[2], Asma Yamani[2]

[1]Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia
[2]Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[3]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, KFUPM, Saudi Arabia
nsalturayeif@iau.edu.sa, {hluqman, g201080740,g201906630}@kfupm.edu.sa

Abstract

Recently, there has been a growing interest in analyzing user-generated text to understand opinions expressed on social media. In NLP, this task is known as stance detection, where the goal is to predict whether the writer is in favor, against, or has no opinion on a given topic. Stance detection is crucial for applications such as sentiment analysis, opinion mining, and social media monitoring, as it helps in capturing the nuanced perspectives of users on various subjects. As part of the ArabicNLP 2024 program, we organized the first shared task on Arabic Stance Detection, `StanceEval 2024`. This initiative aimed to foster advancements in stance detection for the Arabic language, a relatively underrepresented area in Arabic NLP research. This overview paper provides a detailed description of the shared task, covering the dataset, the methodologies used by various teams, and a summary of the results from all participants. We received 28 unique team registrations, and during the testing phase, 16 teams submitted valid entries. The highest classification F-score obtained was 84.38.

1 Introduction

The rapid expansion of social media platforms, online news sources, and digital communication has significantly increased user-generated content in recent years. This surge in online interactions has created a growing need for automated tools and techniques to analyze the opinions and attitudes expressed in these vast text streams effectively. Stance detection, a crucial task in Natural Language Processing (NLP), aims to identify a writer's position or perspective on a specific topic or entity by analyzing their written text and social media activity, including preferences and connections. It has various applications in marketing, politics, and social media analysis.

Stance detection is closely related to sentiment analysis. While sentiment analysis focuses on identifying the explicit sentiment polarity of a text—categorized as Positive, Negative, or Neutral—stance detection classifies the viewpoint of a text towards a specific target as Favor, Against, or None. The target in stance detection is often abstract, like ideological topics, and may not be explicitly mentioned in the text, whereas sentiment analysis usually deals with non-ideological subjects. Additionally, the alignment between sentiment and stance in a text can vary; a text may have a positive sentiment while opposing the target, or vice versa.

In this paper, we summarize the results of the StanceEval 2024[1] shared task where participants were asked to develop models that detect writers' stances (Favor, Against, or None) towards three topics: COVID-19 vaccine, digital transformation, and women empowerment. The MAWQIF dataset, comprising Arabic tweets collected from Twitter, was used for training and evaluating the proposed models. Each tweet in this dataset has been manually annotated with three labels: stance, sentiment, and sarcasm.

2 Literature Review

Stance detection involves classifying a writer's stance on a given subject (target) based on written text (input). The output is typically categorized into one of the following: Favor, Against, None (Alturayeif et al., 2023b). The majority of stance detection research has concentrated on supervised machine learning models developed using manual human annotations (Alturayeif et al., 2023b). This includes approaches using Support Vector Machines (SVM) (Lai et al., 2019; Gómez-Suta et al., 2023), Convolutional Neural Networks (Zhou et al., 2019; Alkhalifa et al., 2021), and Recurrent Neural Networks (Yang et al., 2020). Unsupervised and weakly supervised learning approaches have also

[1]`https://sites.google.com/view/stanceeval/home`

Proceedings of The Second Arabic Natural Language Processing Conference, pages 774–782
August 16, 2024 ©2024 Association for Computational Linguistics

been used for stance detection, including the use of Graph Neural Networks, inferring user's stance based on their past tweets and the tweets of their interaction graph neighbors (Zhang et al., 2023) and label propagation on user interaction networks (Weber et al., 2013). While unsupervised and weakly supervised techniques may not achieve the same level of accuracy as supervised learning methods, they address the challenge of obtaining a sufficient amount of labeled data (Alturayeif et al., 2023b).

Transfer learning is another powerful paradigm utilized for stance detection. It leverages a pre-trained model, initially trained on a broad, general task, and adapts it for the specific target task of stance detection. This approach significantly reduces the labeled data required for the target task, enhancing efficiency and effectiveness in model training (Alturayeif et al., 2023b). Transductive transfer learning was mainly used for domain adaptation (Sun et al., 2022) and cross-lingual learning (Mohtarami et al., 2019). As for inductive transfer learning approaches, sequential transfer learning is the most commonly used type in stance detection, which uses a source task model to improve the target model's performance in a sequential manner (Liu et al., 2022; Ye et al., 2021). Multi-task transfer learning is another type of inductive transfer learning in which several input attributes are learned sequentially or in parallel. In the context of stance detection, these attributes include rumor veracity, sarcasm, and sentiment (Khandelwal, 2021; Ye et al., 2021; Alturayeif et al., 2023a).

With the advent of Large Language Models (LLMs), researchers have started exploring their potential in stance detection (Lan et al., 2024). A recent benchmarking study by Cruickshank et al. (Cruickshank and Ng, 2024) on multiple stance detection datasets (e.g., SemEval 2016 (Mohammad et al., 2016a), COVID-LIES (Hossain et al., 2020), and phemerumors (Kochkina et al., 2018)) demonstrated that few-shot LLM prompting results can be competitive with supervised models. However, these results often show inconsistent performance (Cruickshank and Ng, 2024). Due to the inconsistency, marginal performance differences, and the high inference cost, it remains an open research question whether LLMs should be employed for stance detection (Cruickshank and Ng, 2024).

In this shared task, we invite participants to develop and evaluate stance detection models specifically tailored to the Arabic language. Alternatively, participants can explore various prompting styles of LLMs to perform stance detection on the MAWQIF dataset (Alturayeif et al., 2022), the first dataset designed for Arabic stance detection.

3 Task Description

For the StanceEval-2024 task, participants are invited to showcase their approaches to stance detection. Solutions may utilize machine learning techniques tailored specifically for stance detection or explore alternative methodologies. Participants have the freedom to choose between single-task or Multi-Task Learning (MTL) paradigms. In single-task learning, the focus lies solely on stance data for model development and training. Conversely, MTL-based models offer the flexibility to incorporate additional information, such as sentiment and sarcasm cues from each tweet, to enhance the performance of the stance detection system. The provided dataset encompasses annotations for stance, sentiment, and sarcasm for each tweet.

Specifically, participants are tasked with developing models capable of detecting writers' stances towards three specific topics: COVID-19 vaccine, digital transformation, and women empowerment. The possible stance labels are as follows:

- Favor: Indicates that the author supports the target. This could be explicit support or alignment with the target, or the presence of information such as news, quotes, or stories that reveal support for the target.

- Against: Indicates that the author opposes the target. This could be explicit opposition or alignment with the target, or the presence of information such as news, quotes, or stories that reveal opposition to the target.

- None: Indicates that the tweet provides no hint as to the author's stance toward the target. This could include inquiries or neutral news that does not express any positive or negative position.

Examining the data annotations unveils a disparity between stance and sentiment in texts. It is evident that a tweet can convey a negative sentiment despite holding a favorable stance, and vice versa. Table 1 presents some examples showcasing the annotation of stance, sentiment, and sarcasm dimensions.

Target	Tweet	Stance	Sentiment	Sarcasm
COVID-19 Vaccine	حاشتنا كورونا وطبنا منها ولله الحمد ومانحتاج تطعيم ولاتحسفنا 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 نتمنى 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_مدعيه_النبوه فاهمة تمكين المرأة غلط 😂😂 #Arrest_of_the_prosecutor_of_prophecy she misunderstod women's empowerment 😂😂	None	Neutral	Yes

Table 1: Examples illustrating stance, sentiment, and sarcasm annotations (Alturayeif et al., 2022).

	Train				Test				Total
Target	#Tweets	%Favor	%Against	%None	#Tweets	%Favor	%Against	%None	
COVID-19 Vaccine	1,167	43.62	43.53	12.85	206	43.69	43.69	12.62	1,373
Digital Transformation	1,145	76.77	12.40	10.83	203	76.85	12.32	10.84	1,348
Women Empowerment	1,190	63.87	31.18	4.96	210	63.81	30.95	5.24	1,400
All	3,502	61.34	29.15	9.51	619	61.39	29.08	9.53	4,121

Table 2: Distribution of instances in the Stance Train and Test sets.

3.1 Dataset

The MAWQIF dataset (Alturayeif et al., 2022), a comprehensive resource for Arabic stance detection, is utilized for the StanceEval-2024 shared task. This dataset is meticulously crafted to aid in the analysis and development of models for stance detection, incorporating additional dimensions of sentiment and sarcasm to enhance the performance of stance detection systems.

The annotation process for the MAWQIF dataset involved multiple iterations to ensure high-quality annotations. Initially, annotators encountered challenges when asked to annotate stance, sentiment, and sarcasm simultaneously. To address this, annotation tasks were separated for each dimension, improving the consistency and reliability of the annotations. Each tweet was reviewed by a panel of three to seven annotators. Annotations were considered final when confidence scores exceeded 0.7, or after seven annotations if the required confidence was not met. Test questions were incorporated to evaluate annotator reliability, and annotations from annotators with performance below 80% were excluded from the final dataset.

The dataset comprises a total of 4,121 tweets, covering three distinct topics: "COVID-19 vaccine," "digital transformation," and "women empowerment." Here's the breakdown of tweets for each theme: COVID-19 vaccine: 1,373 tweets, Digital transformation: 1,348 tweets, Women empowerment: 1,400 tweets. This dataset is structured as a multi-label dataset, enabling models to utilize information from various dimensions simultaneously. For the purpose of training and evaluation, we partitioned the dataset into training and testing subsets, with an 85% to 15% split, respectively. Detailed statistics regarding this partitioning can be found in Table 2. A blind test set was provided to ensure fair comparison among participants, and it will be made publicly available after the evaluation period ends.

The dataset, an interactive visualization of the data, and the annotation guidelines can be accessed via the task repository [2].

3.2 Evaluation Metrics

In our evaluation, we utilized the macro F1-score (F_{macro}) as the primary metric. This metric is computed as the average of the F1-scores for the "Favor" and "Against" categories. Specifically, the F_{macro} is calculated separately for each target, and then the overall F_{macro} is computed across all targets. F_{macro} is computed using the following formula:

[2] https://github.com/NoraAlt/Mawqif-Arabic-Stance

$$F_{macro} = \frac{F_{favor} + F_{against}}{2} \quad (1)$$

where F_{favor} and $F_{against}$ are calculated as follows:

$$F_{favor} = \frac{2Precision_{favor}Recall_{favor}}{Precision_{favor} + Recall_{favor}} \quad (2)$$

$$F_{against} = \frac{2Precision_{against}Recall_{against}}{Precision_{against} + Recall_{against}} \quad (3)$$

The F_{macro} evaluation metric offers a comprehensive evaluation of overall performance while addressing imbalanced data. It ensures equal contribution from both majority and minority classes, providing objective results even with imbalanced datasets.

We selected the F_{macro} metric to maintain consistency with other stance detection datasets that report their results using this metric (Mohammad et al., 2016b). It is important to note that the "none" class, although sparse in the data, was still included during training. However, it was not considered in the evaluation because our focus was solely on the "Favor" and "Against" labels for this task. In practice, we view the "None" class as non-interesting or negative in Information Retrieval terms. Misclassifying "None" instances can impact metric scores negatively, and accurately predicting "None" is crucial to avoid labeling penalties. This approach aligns with other stance detection studies, where reporting results specifically for the "favor" and "against" stance labels using F_{macro} is a common practice (Mohammad et al., 2016b; Alturayeif et al., 2023b).

4 Shared Task Teams & Results

4.1 Baselines

Four variants of the BERT model are provided to compare the submitted systems (Alturayeif et al., 2022). We fine-tuned the AraBERT-twitter (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2020), and CAMeLBERT-da (Inoue et al., 2021) models. We will refer to these models as **Baseline I**, **II**, and **III**, respectively.

The baseline models have been fine-tuned using the training data of this shared task. We utilized only the stance label of each training sample during the fine-tuning process. Several pre-processing stages were performed to prepare the training data for fine-tuning pre-trained models. These stages involved removing non-Arabic letters, repeated characters, diacritics, and tatweel. We used a WordPiece tokenizer (Wu et al., 2016) to segment the input text into tokens, and a sequence of up to 128 tokens was fed into BERT-based models. These models were fine-tuned for 20 epochs using the AdamW optimizer with a learning rate of 2e-5.

4.2 Results

In total, we received 28 unique team registrations. During the testing phase, 16 of these teams submitted valid entries. Out of 16 teams, we accepted 13 description papers for publication. Table 3 lists the 16 teams along with the citation of accepted papers.

Table 4 presents the results of the participating teams, ordered based on the F_{macro} metric. We report the results of each topic and the average F_{macro} across all topics. Additionally, we compare these results against our baseline models. As shown in the table, the **AlexUNLP-BH** (Badran et al., 2024) achieved the highest overall F_{macro} score with 84.38% followed by the **MGKM** (Alghaslan and Almutairy, 2024) team with 82.06% F_{macro}, wwhile the **StanceCrafters** (Hasanaath and Alansari, 2024) team secured the third place with 81.68%. Notably, four teams outperformed Baseline-I, while twelve teams surpassed Baseline-II. The gap between Baseline-I and the best-performing model (**AlexUNLP-BH**) is significant at 5.49%, whereas it is 1.52% with the **SMASH** (Hariri and Farha, 2024) model. For Baseline-II, twelve models surpassed it, and even the worst-performing model in this group outperformed it by 0.32 F_{macro}.

4.3 Overview of Participating Systems

AlexUNLP-BH (Badran et al., 2024) implemented target-specific models using AraBERT variants: AraBERTv0.2-Twitter-base for women empowerment and digital transformation topics, and AraBERT COVID-19 for the vaccine topic. To enhance model generalization, they employed data augmentation techniques like random word removal and synonym replacement. Multi-task learning, incorporating sarcasm and sentiment analysis, was utilized to improve stance detection performance. Class imbalances were addressed using weighted cross-entropy loss. Additionally, they

Team	Affiliation
AlexUNLP-BH (Badran et al., 2024)	Alexandria University
BFCAI	Benha University
CUFE (Ibrahim, 2024)	Cairo University
dzStance (Lichouri et al., 2024)	USTHB, CRSTDLA, Algiers 01 University
GITPS	Global IT Professional Services Finland Oy
ISHFMG_TUN (Jaballah, 2024)	University of Tunis, Highsys
ANLP RG (Amal et al., 2024)	University of Sfax
MGKM (Alghaslan and Almutairy, 2024)	King Fahd University of Petroleum and Minerals
PICT (Shukla et al., 2024)	Pune Institute of Computer Technology
Rasid (AlShenaifi et al., 2024)	King Saud University
SMASH (Hariri and Farha, 2024)	University of Edinburgh, University of Sheffield
StanceAlret (Alofi and Mnasri, 2024)	University of Tabuk
StanceCrafters (Hasanaath and Alansari, 2024)	King Fahd University of Petroleum and Minerals
TAO (Melhem et al., 2024)	Palestine Technical University - Kadoorie
Team_Zero (Galal and Kaseb, 2024)	Cairo University
TeamCision	University of Washington

Table 3: List of teams that participated in StanceEval-2024. Teams with accepted papers are cited.

Rank	Team	Women Empowerment	Covid Vaccine	Digital Transformation	Overall F_{macro}
1	AlexUNLP-BH	**88.55**	83.31	**81.27**	**84.38**
2	MGKM	86.96	**84.88**	74.34	82.06
3	StanceCrafters	85.06	79.84	80.14	81.68
4	SMASH	85.03	80.25	75.95	80.41
Baseline-I	AraBERT-twitter	85.77	80.05	70.86	78.89
5	Team_Zero	85.99	73.08	76.8	78.62
6	ANLP RG	83.04	79.38	73.21	78.54
7	PICT	75.52	79.85	78.69	78.02
8	TeamCision	79.49	75.16	77.77	77.48
9	CUFE	83.63	80.38	65.39	76.47
10	Rasid	81.91	71.12	73.94	75.66
11	GITPS	79.34	75.82	69.62	74.93
12	StanceAlret	77.01	70.44	71.95	73.13
Baseline-II	MARBERT	81.64	73.94	62.83	72.81
13	dzStance	74.91	73.43	66.97	71.77
Baseline-III	CAMeLBERT-da	83.96	70.67	59.38	71.34
14	ISHFMG_TUN	73.93	70.19	66.7	70.27
15	TAO	73.3	70.51	64.55	69.45
16	BFCAI	73.2	72.66	50.04	65.3

Table 4: Results on the Blind Test set in F_{macro}.

applied contrastive loss to optimize the distance between similar and dissimilar sentence pairs, further enhancing model performance. The team adopted an ensemble method, combining models trained with different contrastive loss functions, resulting in a comprehensive system. Their approach achieved a macro F1 score of 84.38 and ranked first in the StanceEval 2024 shared task.

ANLP RG (Amal et al., 2024) The CAMeLBERT-da, MARBERT, and AraBERT-twitter models were fine-tuned for sentiment, sarcasm, and stance detection tasks. Preprocessing steps, including HTML tag removal and the replacement of URLs, email addresses, and attributions, were applied to prepare the input text. Among these models, AraBERT-twitter demonstrated the most effective performance in stance detection, achieving the best overall results.

CUFE (Ibrahim, 2024) utilized the newly-released Llama 3-8B model for Arabic stance detection. Each sample in the dataset was prompted using triplets consisting of the topic, text, and stance. The prompt employed the topic and text to predict the stance as a stance detection task. The Llama model was fine-tuned using LoRA. The team achieved a ranking of ninth in the task.

dzStance (Lichouri et al., 2024) employed a combination of Term Frequency-Inverse Document Frequency (TF-IDF) features and Sentence Transformers for stance detection. TF-IDF captures the significance of terms in the document and weights them based on their frequency and rarity across documents. Sentence Transformers encode the contextual and semantic nuances of sentences. These two feature sets are integrated using a weighted fusion strategy and fed into a neural network architecture specifically designed for stance detection.

ISHFMG_TUN (Jaballah, 2024) proposed an ensemble approach for stance detection using a voting technique that incorporated multiple classifiers. During training, features were extracted from the input text and weighted using TF-IDF. Language modeling with different n-grams was employed to extract features at both the character and word levels. These features were then concatenated and fed into a set of classifiers, including SGDClassifier, LinearSVC, Multinomial Naive Bayes, Ridge Classifier, and Random Forest Classifier. Stance prediction was performed using a majority voting technique.

MGKM (Alghaslan and Almutairy, 2024) fine-tuned three large language models (LLMs)—GPT-3.5-Turbo, Meta-Llama-3-8B-Instruct, and Falcon-7B-Instruct—for stance detection using the MAWQIF dataset. Preprocessing involved using the AraBERT pre-processor to handle Arabic tweets with special characters. Models were fine-tuned using a Low-Rank Adaptation (LoRA) approach to manage model sizes locally, with training conducted over three epochs. GPT-3.5-Turbo-0125 emerged as the top-performing model, achieving an impressive F1 score of 82.06 and securing second place in the competition. This study highlights the effectiveness of fine-tuning LLMs for language-specific tasks while acknowledging the computational challenges involved.

PICT (Shukla et al., 2024) employed an MTL approach to create a stance detection system, with stance detection as the primary task and sarcasm and sentiment detection as auxiliary tasks. The model was trained using a focal loss function, and various BERT model variants—including MARBERT, AraBERT, CAMeL-BERT, and QARiB—were evaluated as potential backbones for the system. Among these variants, the MARBERT model achieved the highest reported F1-score when used in the MTL framework.

Rasid (AlShenaifi et al., 2024) employed BERT fine-tuning with a 3-way classification layer as a head, focusing mainly on the MARBERT model due to its inclusion of both dialectal and modern standard Arabic. Additionally, they trained two models based on AraBERT and constructed an ensemble classifier. The Ensemble Classifier comprised Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes, augmented with a TF-IDF Vectorizer. Their experiments demonstrated that MARBERT achieved the best results across all topics, followed by the AraBERT model.

SMASH (Hariri and Farha, 2024) evaluated six BERT-based models (AraBART, AraBERT, mBERT, CAMeLBERT-DA, MARBERT, and QARiB) and seven large language models (LLMs) (AceGPT, Gemma, Llama, Command R, WizardLM, Mistral, Mixtral). BERT and BART-based models were fine-tuned for stance detection using the provided training data, while LLMs were evaluated using a zero-shot setup. The team experimented with different Arabic and English prompts, finding that the best performance was achieved using English prompts and labels. They also reported that using the Arabic translation of the target names improved performance. Notably, Command R and LLAMA 3(70b) models demonstrated higher performance in identifying the stance of Arabic sentences compared to other models, according to the reported results.

StanceCrafters (Hasanaath and Alansari, 2024) proposed an MTL model by leveraging shared knowledge across sentiment, sarcasm, and stance tasks. The model architecture consists of a shared task layer comprising Arabert-Twitter and Camel-Bert BERT-based models combined via an attention mechanism, followed by task-specific layers for each task. Various weighting and aggregation techniques (averaging and attention) were evaluated to find the optimal combination. Additionally, static and relative loss weighting techniques were assessed to give more importance to the primary task. The highest F1-score was obtained by combining two modules via an attention mechanism and assigning a static weight to each task.

TAO (Melhem et al., 2024) fine-tuned ARABERT for stance detection. The preprocessing of the data involved removing non-Arabic letters, numbers, and extra spaces. Additionally, diacritics, Tatweel, and words containing special symbols were eliminated. The cleaned data was then used to fine-tune ARABERT for the stance detection task.

Team_Zero (Galal and Kaseb, 2024) utilized pre-trained language models as feature extractors to avoid the cost of fine-tuning. They employed three Arabic BERT models: MARBERTAV, AraBERTv0.2-Twitter, and Parallel-Sum on Top of BERT (P-SUM-MTL). For MARBERTAV and AraBERTv0.2-Twitter, the team used their pre-trained representations without additional tuning. Features were extracted directly from the models and fed into a logistic regression classifier to determine the stance of each tweet. P-SUM-MTL is a sophisticated architecture that considers embeddings from each of the last four layers of the BERT model. Each layer's output is processed through a subsequent BERT layer, followed by a task-specific classifier in a multitask learning setup. The architecture calculates the overall loss by summing the losses from all individual classifiers, thereby effectively integrating insights from multiple tasks. The team applied majority voting for the outputs of the three models to obtain the final results.

StanceAlert (Alofi and Mnasri, 2024) employed comparative analysis to correlate dates with topics. For example, they observed that discussions on women empowerment were prevalent from 2020 until September 2021, followed by a decline. They then utilized a histogram-based approach to identify the most frequent words for each topic and stance pair. For fine-tuning, they used the bert-base-arabertv02-twitter model with logistic regression to predict stance. Through this approach, the authors achieved 12th place in the task.

5 Conclusion

In conclusion, the StanceEval shared task, organized as part of the ArabicNLP 2024 program, has made significant strides in advancing Arabic stance detection. With the participation of 28 teams, 16 of which submitted valid entries and 13 provided detailed description papers, it reflects the growing interest and engagement in this research area. The comprehensive analysis presented in this paper underscores the diverse approaches and methodologies employed by the teams, offering valuable insights into the current state of stance detection in Arabic texts. The results and comparative analysis not only highlight the strengths and limitations of various models but also lay the groundwork for future improvements and innovations. This shared task represents a crucial step towards the development of more effective and robust tools for analyzing user-generated text, with broad applications across politics, marketing, and social media analysis.

Acknowledgments

The authors would like to acknowledge the support received from the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant no. JRC-AI-RFP-14.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Mamoun Alghaslan and Khaled Almutairy. 2024. Mgkm at stanceeval2024: Fine-tuning large language models for arabic stance detection. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2021. Opinions are made to be changed: Temporally adaptive stance classification. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, HT '21. ACM.

Eman Sweiaad S. Alofi and Sami Mnasri. 2024. Stancealert: Effective logistic regression for stance detection. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Nouf AlShenaifi, Nourah Alangari, and Hadeel Al-Negheimish. 2024. Rasid at stanceeval: Fine-tuning marbert for arabic stance detection. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023a. Enhancing stance detection through sequential weighted multi-task learning.

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

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.

Mezghani Amal, Rahma Boujelbane, and Mariem Ellouze. 2024. Anlp rg at stanceeval2024: Comparative evaluation of stance, sentiment and sarcasm detection. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv*.

Mohamed Badran, Mo'men Hamdy, Marwan Torki, and Nagwa El-Makky. 2024. Alexunlp-bh at stanceeval2024: Multiple contrastive losses ensemble strategy with multi-task learning for stance detection in arabic. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Iain J. Cruickshank and Lynnette Hui Xian Ng. 2024. Prompting and fine-tuning open-sourced large language models for stance classification. *Preprint*, arXiv:2309.13734.

Omar Galal and Abdelrahman Kaseb. 2024. Team_zero at stanceeval2024: Frozen plms for arabic stance detection. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Manuela Gómez-Suta, Julián Echeverry-Correa, and José A. Soto-Mejía. 2023. Stance detection in tweets: A topic modeling approach supporting explainability. *Expert Systems with Applications*, 214:119046.

Youssef Al Hariri and Ibrahim Abu Farha. 2024. Smash at stanceeval 2024: Prompt engineering llms for arabic stance detection. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Ahmed Abul Hasanaath and Aisha Alansari. 2024. Stancecrafters at stanceeval2024: Multi-task stance detection using bert ensemble with attention based aggregation. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Michael Ibrahim. 2024. Cufe at stanceeval2024: Arabic stance detection with fine-tuned llama-3 model. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

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. *arXiv preprint arXiv:2103.06678*.

Mustapha Jaballah. 2024. Ishfmg_tun at stanceeval: Ensemble method for arabic stance evaluation system. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Anant Khandelwal. 2021. Fine-tune longformer for jointly predicting rumor stance and veracity. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, CODS COMAD 2021. ACM.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. *Preprint*, arXiv:2310.10467.

Mohamed Lichouri, Khaled Lounnas, OUARAS Khelil Rafik, Mohamed ABi, and Anis Guechtouli. 2024. dzstance@stanceeval2024: Arabic stance detection based on sentence transformers. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. *Preprint*, arXiv:2205.00619.

Anas Melhem, Osama Hamed, and Thaer Sammar. 2024. Tao at stanceeval2024 shared task: Arabic stance detection using arabert. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

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

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

Mitra Mohtarami, James Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4442–4452, Hong Kong, China. Association for Computational Linguistics.

Ishaan Shukla, Ankit Vaidya, and Geetanjali Vinayak Kale. 2024. Pict at stanceeval2024: Stance detection in arabic using ensemble of large language models. In *Proceedings of the ArabicNLP 2024*, Bangkok, Thailand (Hybrid).

Qingying Sun, Xuefeng Xi, Jiajun Sun, Zhongqing Wang, and Huiyan Xu. 2022. Stance detection with a multi-target adversarial attention network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–21.

Ingmar Weber, Venkata R. Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13. ACM.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yuanyu Yang, Bin Wu, Kai Zhao, and Wenying Guo. 2020. Tweet stance detection: A two-stage dc-bilstm model based on semantic attention. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*. IEEE.

Kai Ye, Yangheran Piao, Kun Zhao, and Xiaohui Cui. 2021. *Graph Enhanced BERT for Stance-Aware Rumor Verification on Social Media*, page 422–435. Springer International Publishing.

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. 2023. Doubleh: Twitter user stance detection via bipartite graph neural networks. *Preprint*, arXiv:2301.08774.

Shengping Zhou, Junjie Lin, Lianzhi Tan, and Xin Liu. 2019. Condensed convolution neural network by attention over self-attention for stance detection in twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.