Mawqif: A Multi-label Arabic Dataset for Target-specific Stance Detection

Nora Alturayeif[1,2], **Hamzah Luqman**[1,3], **and Moataz Ahmed**[1,4]

[1]King Fahd University of Petroleum and Minerals, Saudi Arabia
[2]Imam Abdulrahman Bin Faisal University, Saudi Arabia
[3]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, KFUPM
[4]Interdisciplinary Research Center of Intelligent Secure Systems (IRC-ISS), KFUPM
[1]{g201902190,hluqman,moataz}@kfupm.edu.sa

Abstract

Social media platforms are becoming inherent parts of people's daily life to express opinions and stances toward topics of varying polarities. Stance detection determines the viewpoint expressed in a text toward a target. While communication on social media (e.g., Twitter) takes place in more than 40 languages, the majority of stance detection research has been focused on English. Although some efforts have recently been made to develop stance detection datasets in other languages, no similar efforts seem to have considered the Arabic language. In this paper, we present Mawqif, the first Arabic dataset for target-specific stance detection, composed of 4,121 tweets annotated with stance, sentiment, and sarcasm polarities. Mawqif, as a multi-label dataset, can provide more opportunities for studying the interaction between different opinion dimensions and evaluating a multi-task model. We provide a detailed description of the dataset, present an analysis of the produced annotation, and evaluate four BERT-based models on it. Our best model achieves a macro-F_1 of 78.89%, which shows that there is ample room for improvement on this challenging task. We publicly release our dataset, the annotation guidelines, and the code of the experiments.[1]

1 Introduction

Currently, online forums and social media platforms are being inherent parts of people's daily life as a media of expressing their stances toward different targets (e.g., events, politics, services, or controversial news). Consequently, the demand for automatic solutions for stance detection significantly increases as the volume of unstructured data does.

Stance detection is the task of predicting whether the author of a written text is in favor of, against, or neutral toward a subject of interest (i.e., target), in which the stance is explicitly or implicitly stated in the text (Küçük and Fazli, 2020; AlDayel and Magdy, 2021). Automatic and high-performance solutions for stance detection can play a valuable role in decision-making for politicians, businesses, and authorities. The input to the stance detector is usually a pair of written text and a target. However, other inputs can be used to boost the model performance such as the user's social activity on the social media platforms (e.g., retweets and likes).

Existing stance detection datasets can be categorized based on the target dependency into *target-specific*, *cross-target*, and *target-independent*. In *target-specific* stance detection, a specific target (e.g., Donald Trump or BREXIT referendum) has to be given along with the user's text, and sometimes the user's information, in order to detect the stance toward the predefined target. In *cross-target* stance detection, the objective is to build a classifier that can transfer the learned knowledge between targets using a large dataset that comprise a wider range of different targets. In the *target-specific* and *cross-target* tasks, the target of the stance is an explicit entity (e.g., person, event, or controversial issue), whereas the target in *target-independent* tasks is a claim or a piece of fake news and the objective is to detect whether the comments are confirming the claim/news or denying its veracity.

A significant number of stance detection techniques have been proposed in the literature. However, most of these studies used an old public dataset, SemEval-2016 (Mohammad et al., 2016), including those published recently (Chen et al., 2021; Li et al., 2021b; Al-Ghadir et al., 2021; Allaway et al., 2021; Liang et al., 2021). We believe that more benchmarked stance detection datasets should be released under a common open license for public usage. Non-English data, multilingual data, and annotations of other opinion dimensions (e.g., sarcasm and emotions) should all be considered for establishing new stance detection datasets.

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

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 174 - 184
December 8, 2022 ©2022 Association for Computational Linguistics

We aim to facilitate the research on target-specific stance detection of Arabic micro-blogs. To our knowledge, this problem has not been studied for the Arabic language and there is no publicly available dataset for Arabic that can be used for target-specific stance detection. Arabic is a challenging language for most natural language processing (NLP) applications due to its unique nature in the variety of dialectics and its rich and complex morphology (Badaro et al., 2020). Furthermore, different from media that use Modern Standard Arabic (MSA) with formal linguistic criteria, social media texts represent dialectal Arabic and contain an informal writing style (e.g., spelling errors, abbreviations, irregular grammar, emojis, and symbols). Thus, automatically detecting the user's stance on social media, specifically in Arabic, is a worthwhile and challenging task. In addition, the increase of Arabic content on social media, and the mobilized masses for political and economic changes in the Middle East have motivated us to search in this direction.

In this paper, we release MAWQIF, the first Arabic dataset that can be used for target-specific stance detection. This dataset consists of 4,121 tweets in multi-dialectal Arabic. Each tweet is annotated with a stance toward one of three targets: "COVID-19 vaccine," "digital transformation," and "women empowerment." In addition, this is a multi-label dataset where each data point is annotated for stance, sentiment, and sarcasm, which will provide a benchmark for the three tasks. It will also help in analyzing the interaction between the different opinion dimensions (i.e., stance, sentiment, and sarcasm).

Our contributions in this paper can, therefore, be summarized as follows. **1)** We construct and release MAWQIF, the first multi-label Arabic dataset for stance detection. The proposed dataset consists of 4,121 tweets covering three topics (i.e., targets) that are controversial in the Middle East. We also provide a detailed description of the dataset and an analysis of the produced annotation; **2)** The proposed dataset is annotated for stance, sentiment, and sarcasm. This provides more opportunities for studying the interaction between different opinion dimensions, and evaluating a model trained on different opinion dimensions in a multi-task paradigm to boost the performance of stance detection; **3)** We benchmark the proposed dataset on the stance detection task and evaluate the performance of four BERT-based models.

2 Related work

Stance detection is a relatively new field of study; however, considerable effort has been devoted into building datasets for stance detection tasks. From the definitions of the three stance detection tasks (presented in Section 1); the structure of the datasets used for target-independent tasks is different than the datasets used for target-specific or cross-target tasks. In target-independent stance detection, each input entry is usually in the form of a pair of textual claims and responses. Examples of target-independent datasets are: Emergent (Ferreira and Vlachos, 2016), IBM Debater (Bar-Haim et al., 2017), Pheme (Kochkina et al., 2017), RumourEval-17 (Derczynski et al., 2017), FNC-1 (Hanselowski et al., 2018), Args.me (Ajjour et al., 2019), Perspectrum (Chen et al., 2019), RumourEval-19 (Gorrell et al., 2019), Arabic News Stance (Khouja, 2020), and (Baly et al., 2018). Meanwhile, the input entry for target-specific and cross-target stance detection systems usually consists of a text and target pair.

Several datasets have been proposed for target-specific and cross-target stance detection. These datasets have been collected from different platforms such as social media (Mohammad et al., 2016; Xu et al., 2016; Sobhani et al., 2017; Taulé et al., 2017; Küçük and Can, 2018; Lai et al., 2018; Conforti et al., 2020; Lai et al., 2020; Cignarella et al., 2020; Grimminger and Klinger, 2021; Zotova et al., 2021), debate websites (Stab et al., 2018; Hosseinia et al., 2020; Vamvas and Sennrich, 2020), and news commentaries (Hercig et al., 2017; Allaway and Mckeown, 2020). With regard to language orientation, most of the available stance detection datasets are monolingual where their data are available in one language. The majority of these monolingual datasets are in English language (Mohammad et al., 2016; Sobhani et al., 2017; Stab et al., 2018; Allaway and Mckeown, 2020; Conforti et al., 2020; Lai et al., 2020; Hosseinia et al., 2020; Grimminger and Klinger, 2021). For Italian, Lai et al. (2018) and Cignarella et al. (2020) collected tweets targeting the Italian constitutional reform and the Sardines movement, respectively. Similarly, Küçük and Can (2018) collected Turkish tweets targeting football clubs. Furthermore, a dataset for Chinese language is presented in (Xu et al., 2016), and a Czech stance detection dataset is presented

Language	Dataset Name / Ref.	Targets	Annotation	Size
English	SemEval-2016 Task 6 (Mohammad et al., 2016)	Atheism, Climate change, Feminist movement, Hillary Clinton, Abortion legalization	Stance, Sentiment	4,163 Tweets
	Multi-target SD (Sobhani et al., 2017)	2016 US presidential electors	Stance	4,455 Tweets
	UKP (Stab et al., 2018)	8 controversial topics	Stance	25,492 Comments
	Procon20 (Hosseinia et al., 2020)	419 controversial issues	Stance	6,094 Comments
	VAST (Allaway and Mckeown, 2020)	Several topics	Stance	23,525 Comments
	WT-WT (Conforti et al., 2020)	Health insurance companies	Stance	51,284 Tweets
	TW-BREXIT (Lai et al., 2020)	BREXIT referendum	Stance	1,800 Triplets of tweets
	Election-2020 (Grimminger and Klinger, 2021)	2020 US presidential electors	Stance, Hate speech	3,000 Tweets
Italian	ConRef-STANCE-ita (Lai et al., 2018)	Italian constitutional reforms	Stance	963 Triplets (tweet, retweet, reply)
	SardiStance (Cignarella et al., 2020)	Sardines movement	Stance	3,242 Tweets
Chinese	NLPCC-2016 Task 4 (Xu et al., 2016)	5 topics	Stance	3,250 Weibo posts
Czech	Hercig et al. (2017)	Miloš Zeman, Smoking ban	Stance, Sentiment	5,423 Comments
Turkish	Küçük and Can (2018)	Football clubs	Stance	1,065 Tweets
Spanish, Catalan	IberEval 2017 (Taulé et al., 2017)	Catalan independence	Stance	5,400 Tweets (for each language)
	Zotova et al. (2021)	Catalan independence	Stance (automatic annotation)	Spanish: 10K Tweets, Catalan: 10K Tweets
German, French, Italian	X-stance (Vamvas and Sennrich, 2020)	150 political issues	Stance (automatic annotation)	German: 40,200, French: 14,129, Italy: 1,173

Table 1: Publicly available datasets for target-specific and cross-target stance detection.

in (Hercig et al., 2017). However, few datasets are multilingual where more than one language is considered in collecting the data. Vamvas and Sennrich (2020) proposed a multilingual dataset with French, German, and Italian languages. Two other datasets considered Catalan and Spanish languages in one dataset (Taulé et al., 2017; Zotova et al., 2021). Table 1 summarizes the publicly available datasets used for target-specific and cross-target stance detection.

In our dataset, we attempt to address two gaps; the language and the annotation of other opinion dimensions. Despite the growing interest in studying stance detection, no study, as far as we know, considered Arabic language for target-specific stance detection. In this paper, we release the first Arabic target-specific stance detection dataset. It is worthwhile noting that there are two stance detection datasets that target Arabic language (Khouja, 2020; Alhindi et al., 2021). However, these two datasets are dedicated to study claim verification, as they consist of claim/reference pairs to predict the stance of a claim toward the reference sentence. Thus, they cannot be used for building a target-specific stance detection model. In addition, the two datasets are comprising texts in modern standard Arabic, which is not the language used in social media debates where dialectal Arabic is quite prevalent.

Moreover, most of the existing datasets annotated each text with stance labels (Favor, Against, None). Other studies considered the sentiment polarity during data annotation. The aim of involving sentiment annotation was to analyze the interaction between stance and sentiment in order to boost the performance of stance detection (Mohammad et al., 2016; Hosseinia et al., 2020). However, there is no study to the best of our knowledge has considered sarcasm features for stance detection. According to the findings of a comparative empirical study by (Ghosh et al., 2019), the main source of misclassification in stance detection is texts with sarcastic content. Therefore, studying sarcasm could be beneficial for improving the performance of stance detection models. We thus proposed to annotate our dataset with sarcasm in addition to stance and sentiment polarities. Our dataset is established in order to create a novel Arabic linguistic resource for stance, sentiment, and sarcasm.

3 Mawqif Dataset

In this section, we explain the procedure followed to collect a set of opinions (texts) toward selected targets for stance detection. We also present the crowdsourcing setup used for stance annotation and discuss the statistics of the proposed dataset.

3.1 Data Collection and Filtering

Most of the available stance detection datasets focus mainly on a narrow range of political topics, such as elections and referendums. In contrast, we extended the considered domains in our dataset to include other topics related to hot social issues in the Middle East. Similar to prior works (Li et al., 2021a; Conforti et al., 2020; Lai et al., 2020; Sobhani et al., 2016; Mohammad et al., 2016) that targeted multiple topics, we considered three targets: "COVID-19 vaccine," "digital transformation," and "women empowerment." The proposed dataset has been collected from Twitter platform. We crawled tweets using Snscrape[2] crawler which is a python library for social networking services.

A set of keywords and query hashtags were used as seeds to collect target-related tweets. This phase resulted in collecting around 400K tweets. It should be noted that a considerable number of collected tweets contain stance-indicative hashtags; however, this does not imply that the tweet will take the same stance as indicated by the hashtag. An example from our dataset:

#لا_للتطعيم_الاجباري لو تطعيم كورونا مضر كان حتى
تطعيمات الأطفال مضره، توكل على الله وطعم

#No_to_compulsory_vaccination If the corona vaccine is harmful then even the vaccines for children are harmful, so put your trust in Allah and get it

The second phase in the data collection stage was to filter and prepare the collected data. We performed the following preprocessing steps: 1) We kept only the Arabic tweets, which include multi dialects, and removed tweets in other languages. 2) We removed duplicates and retweets. 3) Tweets from news media accounts were eliminated using the information contained in *user_description* attribute available in the Snscrape tweet object. 4) We defined a set of keywords and phrases that usually appear in advertisements and adult tweets to exclude these types of tweets. 5) Tweets were cleaned from URLs and user mentions. Applying these filters resulted in reducing the collected tweets to around 200K tweets for all three targets combined. Finally, we randomly sampled around 1,400 tweets for each target, obtaining 4,121 tweets in total for annotation.

3.2 Annotation

To annotate our data, we used Appen crowdsourcing platform[3] to hire native Arabic speakers who live in Arab countries for the annotation task. We asked the contributors (i.e., annotators) to perform stance, sentiment, and sarcasm annotations for each tweet of the proposed dataset. This will help in using the dataset for these three tasks.

To build our quality control step, we conducted the annotation process in multiple iterations. In each iteration, we used a batch of 100 tweets for evaluating annotation quality. Initially, we created an annotation form that provides instructions for annotating the three dimensions (i.e., stance, sentiment, and sarcasm), and asked the annotators to annotate each tweet with the three dimensions at the same time. We noticed that the assignment was quite challenging, resulting in a low score of inter-agreement between annotators. Therefore, we designed a separate annotation form for each dimension (i.e., we assigned three separate tasks for different annotators). We noticed that letting the annotator focus on one task at a time was much easier and resulted in a higher inter-agreement between the annotators. In addition, it resulted in greater consensus among the annotators. Therefore, rather than generating a single annotation form for all three dimensions, we picked the latter approach for our annotation process.

In the stance annotation form, we asked the annotators to read a tweet and identify its stance (i.e, Favor, Against, None) toward a predefined target. The annotators were also asked to determine if the target is mentioned explicitly or implicitly in the

[2] https://github.com/JustAnotherArchivist/snscrape

[3] https://appen.com

tweet. We designed similar annotation forms to determine the sentiment of a tweet (i.e, Positive, Negative, or Neutral), and to determine if the tweet contains sarcastic content or not. With regard to sarcasm, we define it, according to the Cambridge English dictionary, as: "*Sarcastic means the text expresses an evaluation whose literal polarity is different from the intended polarity to hurt someone emotionally or criticize something in a humorous way*". To ensure the consistency between the annotation of the proposed dataset and other similar datasets, we followed the stance and sentiment annotation guidelines formulated in (Mohammad et al., 2017). Our dataset release is accompanied by the annotation guidelines.

Each tweet–target pair was annotated by three to seven annotators. We require to stop collecting annotations on a row when the row's confidence score is above 0.7 or when a maximum of seven annotations is reached. Appen system provides a mechanism to compute the confidence score based on the level of agreement among multiple annotators, weighted by the trust scores of the annotators. We control the quality of the annotation by 420 test questions with correct labels for stance, sentiment, and sarcasm that were interleaved between the regular questions. An annotator's trust score was computed on these test questions; under-performers who got scores below 80% were eliminated and all their submitted annotations were also ignored.

3.3 Dataset Statistics

The distribution of the confidence in the annotations of the three dimensions (i.e., stance, sentiment, and sarcasm) is shown in Figure 1. Based on our analysis in evaluating the annotation quality using our test questions, the confidence threshold for high-confidence annotation was set to 0.7. We observed a lower inter-agreement on the sentiment annotation, with around 30% of annotations' confidence score below 0.7 (light red in Figure 1). This, in line with our beliefs, confirm the highly subjective nature of sentiment annotation. Meanwhile, stance annotations produced a higher agreement, with 15% were considered as low-confidence. The highest confidence annotations were achieved in sarcasm, with only 5.75% below 0.7 score.

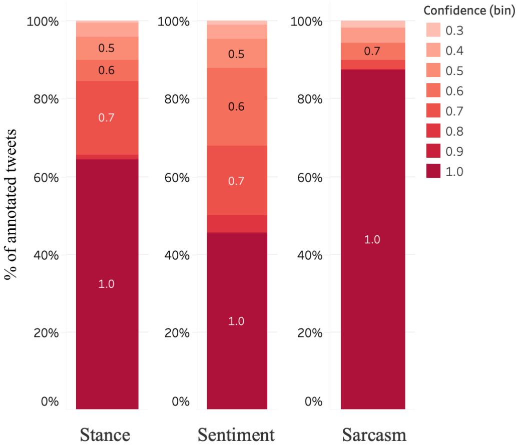


Figure 1: Distributions of the confidence in the stance, sentiment, and sarcasm annotations.

The MAWQIF dataset contains 4,121 annotated tweets representing three targets: "COVID-19 vaccine" with 1,373 tweets, "digital transformation" with 1,348 tweets, and "women empowerment" with 1,400 tweets. This dataset is a multi-label dataset where each tweet is annotated for stance, sentiment, and sarcasm. Table 2 show some examples from MAWQIF dataset. We split the dataset into training and testing sets with 85% and 15%, respectively. The data split statistics are shown in Table 3.

Figure 2 illustrates the labels' distribution across all targets, and the distribution per target. As observed from this figure, the percentage of tweets that do not have a clear stance and are labeled as *none* are low (9.51%) compared to the ones labeled as *neutral* sentiment (31%). This demonstrates that neutral tweets do not imply that they do not show any stance. Regarding sarcasm, most of the tweets were annotated as non-sarcasm (95.39%). This is expected, given that we were not targeting sarcastic text in our dataset.

The labels' distribution varies between the three targets. Tweets discussing digital transformation tend to lean toward a favorable stance compared to the other targets. Regarding sentiment polarity, positive content appears more frequently when discussing women empowerment or digital transformation, compared to the COVID-19 vaccine topic with only 25% positive tweets. Furthermore, sarcastic content appears more frequently in COVID-19 vaccine related tweets.

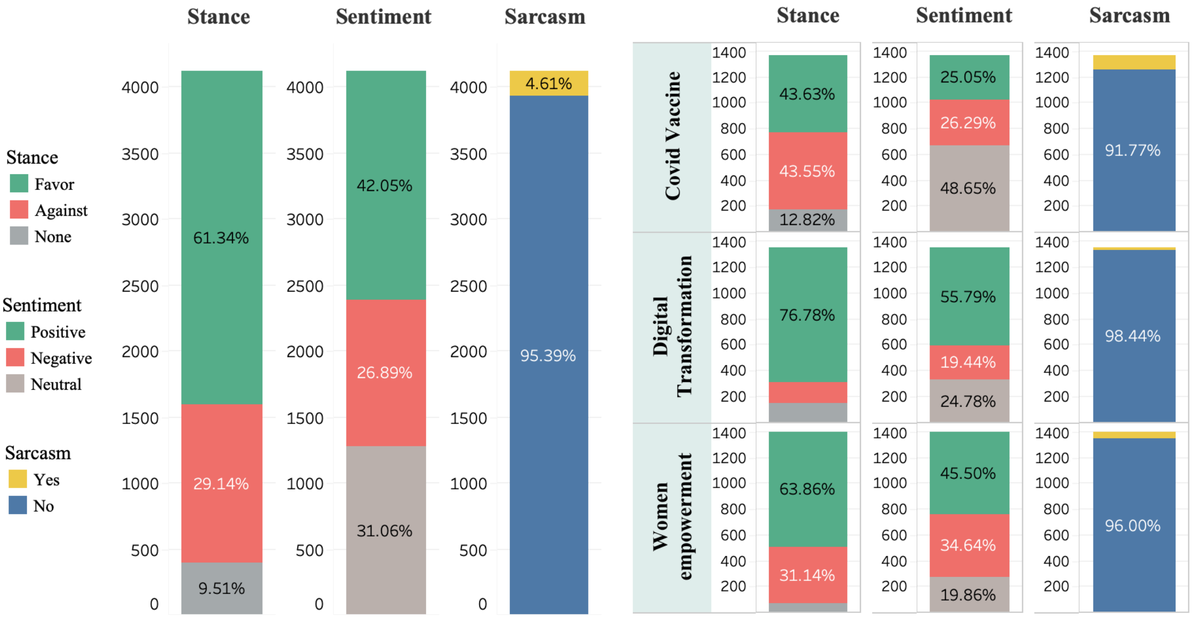
We also studied the association between stance and sentiment, and between stance and sarcasm through a co-occurrence heatmap (Figure 3). Examination of the stance-sentiment matrix reveals that stance is not always aligned with the sentiment for a target within a text. This implies that a tweet may have a negative polarity, but the stance is in favor, or vice versa (some examples are shown in

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 2: Examples from MAWQIF dataset that show how stance may not align with sentiment polarity.

	Train				Test				Total
Target	#Tweets	%Favor	%Against	%None	#Tweets	%Favor	%Against	%None	
COVID-19 Vaccine	1167	43.62	43.53	12.85	206	43.69	43.69	12.62	1373
Digital Transformation	1145	76.77	12.40	10.83	203	76.85	12.32	10.84	1348
Women Empowerment	1190	63.87	31.18	4.96	210	63.81	30.95	5.24	1400
All	3502	61.34	29.15	9.51	619	61.39	29.08	9.53	4121

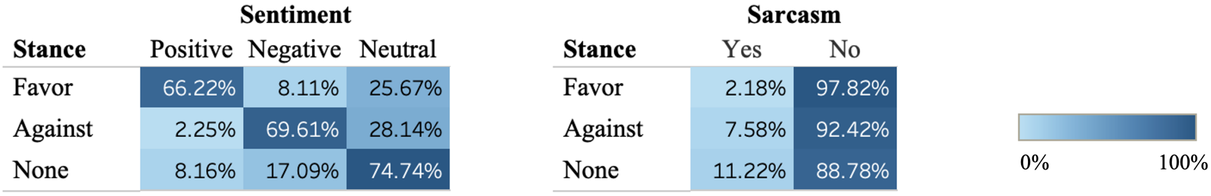
Table 3: Data split statistics of MAWQIF dataset.



(a) Overall labels' distribution

(b) Labels' distribution per target

Figure 2: Labels' distribution in MAWQIF dataset.



	Sentiment		
Stance	Positive	Negative	Neutral
Favor	66.22%	8.11%	25.67%
Against	2.25%	69.61%	28.14%
None	8.16%	17.09%	74.74%

	Sarcasm	
Stance	Yes	No
Favor	2.18%	97.82%
Against	7.58%	92.42%
None	11.22%	88.78%

Figure 3: Association between Stance and Sentiment, (a) Stance-Sentiment association, (b) Stance-Sarcasm association.

Table 2). Around 34% of favor tweets are actually not positive, and 31% of tweets with negative stances are annotated with a non-negative sentiment. From the stance-sarcasm matrix, we can observe that sarcastic content appears more in instances that are labeled as *against* compared to instances of favorable stance.

4 Benchmark Experiments

In this section, we present benchmarking experiments performed on the target-specific stance detection task. As mentioned earlier, the main purpose of MAWQIF dataset is stance detection. Therefore, we considered only the stance detection task for the benchmark experiments. However, the sentiment and sarcasm annotations could be used in further experiments (i.e, future studies) to analyze the interaction between the three dimensions.

Models BERT-based models have been shown to be effective in a variety of text classification tasks (González-Carvajal and Garrido-Merchán, 2020), including dialectical Arabic text (Alturayeif and Luqman, 2021). Thus, we chose to develop a BERT-based classifier that we fine-tuned for target-specific stance detection. Specifically, we fine-tuned the following four BERT-based models for stance detection:

1. CAMeLBERT-da, is a BERT-based model trained on 5.8 billion tokens from the Dialectal Arabic (DA) dataset (Inoue et al., 2021).
2. MARBERT, is a BERT-based model trained on 15.6 billion tokens from 1 billion Arabic tweets (Abdul-Mageed et al., 2020).
3. AraBERT, is trained on 8.6 billion tokens from five datasets consisting of Modern Standard Arabic (MSA) text (Antoun et al., 2020).
4. AraBERT-twitter, is trained by extending the training of AraBERT (v0.2) on 60 million Arabic tweets (Antoun et al., 2020).

We fine-tuned the four pre-trained models and built a standard pipeline under the PyTorch Lightning framework. The fine-tuning code is available online along with our dataset. The proposed system starts by preprocessing the Arabic texts by removing diacritics, tatweel, non-Arabic letters, and repeated characters. Then, a WordPiece (Wu et al., 2016) tokenizer is used to split the input text into tokens compatible with BERT-based models. For classification, the hidden representation of the [CLS] token is fed into a feed-forward layer along with a Softmax function. We set the maximum sequence length to 128 tokens, and the batch size to 32. Each of the four models is fine-tuned for 20 epochs; AdamW optimizer (Loshchilov and Hutter, 2017) is used with a learning rate of 2e-5. The hyper-parameters used in these experiments have been selected empirically.

Evaluation Metrics We evaluated our baseline models using F_{avg2} and F_{avg3} scores. F_{avg2} is the macro-average F1 over the "favor" and "against" stance labels (the "none" class was ignored since it was scarcely in the data). This score is computed as follows:

$$F_{avg2} = \frac{F_{favor} + F_{against}}{2} \quad (1)$$

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

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

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

We selected F_{avg2} metric to align with other stance detection datasets that report their results using F_{avg2} metric (Mohammad et al., 2016). We are also reporting our results using F_{avg3} that considers all stances and it is computed as follows:

$$F_{avg3} = \frac{F_{none} + F_{favor} + F_{against}}{3} \quad (4)$$

Results Tables 4 and 5 present the obtained results of the proposed models with the development and test sets, respectively. The development set was obtained by dividing the training set into 5-folds and training the model with cross-validation. As shown in Tables 4 and 5, AraBERT-twitter model yields the best overall and per-target performance. This can be attributed to the type of the train data (i.e, dialectical Arabic tweets) that were used to train AraBERT-twitter model, which is similar to the type of Arabic tweets used in MAWQIF dataset. Furthermore, we can observe that the best performed model (i.e. AraBERT-twitter) and the other three models (CAMeLBERT-da, MARBERT, and AraBERT) generalized quite well to the test data, even achieving higher accuracies and macro-F_1 scores.

Although MARBERT was trained on dialectical Arabic tweets, its performance is low compared

	COVID-19 Vaccine		Digital Transformation		Women Empowerment		Overall					
Model	F_{avg2}	F_{avg3}	F_{avg2}	F_{avg3}	F_{avg2}	F_{avg3}	F_{favor}	$F_{against}$	F_{none}	F_{avg2}	F_{avg3}	Acc
CAMeLBERT-da	71.84	57.42	59.36	42.35	73.61	49.07	79.90	56.63	12.30	68.27	49.61	71.72
MARBERT	73.94	**63.96**	49.30	44.99	78.31	52.21	82.83	51.53	**26.79**	67.18	53.72	74.86
AraBERT	76.01	57.62	59.51	49.19	73.41	48.94	80.85	58.44	16.47	69.64	51.92	73.77
AraBERT-twitter	**76.77**	61.71	**62.25**	**56.31**	**84.91**	**56.60**	**83.78**	**65.51**	25.34	**74.64**	**58.21**	**76.56**

Table 4: Stance detection results on the development set.

	COVID-19 Vaccine		Digital Transformation		Women Empowerment		Overall					
Model	F_{avg2}	F_{avg3}	F_{avg2}	F_{avg3}	F_{avg2}	F_{avg3}	F_{favor}	$F_{against}$	F_{none}	F_{avg2}	F_{avg3}	Acc
CAMeLBERT-da	70.67	59.61	59.38	47.28	83.96	55.97	81.78	60.90	20.19	71.34	54.29	73.61
MARBERT	73.94	63.96	62.83	50.77	81.64	**59.98**	82.91	62.70	**29.11**	72.81	58.24	75.97
AraBERT	73.39	62.26	67.43	52.36	78.09	52.06	82.17	63.77	20.74	72.97	55.56	75.10
AraBERT-twitter	**80.05**	**65.49**	**70.86**	**63.03**	**85.77**	57.18	**86.54**	**71.25**	27.91	**78.89**	**61.90**	**79.78**

Table 5: Stance detection results on the test set.

to AraBERT-twitter. This may be explained by the fact that MARBERT was trained with masked-language modeling (MLM) objective only, whereas AraBERT was trained with both MLM and the next sentence prediction (NSP) objectives. While MLM aims to capture the relationship between words, NSP aims to understand longer-term dependencies between sentences. Thus, NSP objective could improve the ability to capture more information in the sentence–stance pairs that appear in our training dataset.

CAMeLBERT-da was trained on dialectical Arabic data collected from social media sites and other resources. However, CAMeLBERT-da has a lower performance due to the smaller size of its training data compared to the data used to train AraBERT-twitter. CAMeLBERT-da was trained on 5.8 billion words with a vocabulary size of 30K, while AraBERT-twitter was trained on 8.6 billion words with a vocabulary size of 60K in addition to 60M multi-dialect tweets.

It is also noticeable in the obtained results that the performance of all models in detecting the *none* stance is low compared with other stances. This can be attributed to the small number of tweets with *none* stance used in model training. However, *none* is a class that is not of interest as the ultimate goal is to infer if the author of a written text is in favor of or against a specific target. On other hand, the obtained results with the *favor* stance were high compared with the *against* stance in all experimented models. This indicates that there is room for improvement in all models, where a model can benefit from the techniques that mitigate the impact of class imbalance.

Furthermore, we can observe from Table 5 that the performance scores of all models were the highest with the “women empowerment” target. This might be an indication of strong signals appearing in the tweets discussing women empowerment that separate instances that are in favor and those that are against.

5 Conclusion

We introduced Mawqif, the first multi-label Arabic dataset for target-specific stance detection. The proposed dataset consists of 4,121 multi-dialectal Arabic tweets targeting three topics that are controversial in the Middle East. Mawqif is not limited to stance annotation, it is further annotated with sentiment and sarcasm polarity. Thus, Mawqif can serve as a new benchmark for three tasks: stance detection, sentiment analysis, and sarcasm detection. In addition, it can enable future research in studying the interaction between different opinion dimensions, and evaluating multi-task models. We also presented a detailed description of the dataset and an analysis of the produced annotation. Lastly, we experimented on the target-specific stance detection task and establish strong baselines based on four BERT-based models.

Future work may improve upon the reported results by minimizing the effects of class imbalance, which can be accomplished by oversampling or undersampling techniques, or by training with weighted loss. Another interesting direction for further research is developing a joint neural architecture based on a multi-task learning paradigm that jointly models sentiment and sarcasm to boost the performance of stance detection.

To facilitate future research, we publicly release our dataset, the annotation guidelines, and the code that can be used to reproduce the presented evaluation results.

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 JRC-AI-RFP-05.

References

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

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. volume 11793 LNAI of *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 48–59.

Abdulrahman I. Al-Ghadir, Aqil M. Azmi, and Amir Hussain. 2021. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion*, 67:29–40.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: state of the art and trends. *Information Processing and Management*, 58.

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. pages 57–65.

Emily Allaway and Kathleen Mckeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8913–8931.

Emily Allaway, Malavika Srikanth, and Kathleen Mckeown. 2021. Adversarial learning for zero-shot stance detection on social media. pages 4756–4767.

Nora Alturayeif and Hamzah Luqman. 2021. Fine-grained sentiment analysis of arabic covid-19 tweets using bert-based transformers and dynamically weighted loss function. *Applied Sciences*, 11(22):10694.

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

Gilbert Badaro, Ramy Baly, Hazem Hajj, Wassim El-hajj, Khaled Bashir Shaban, Nizar Habash, Ahmad Al-sallab, and A L I Hamdi. 2020. A survey of opinion mining in arabic : A comprehensive system perspective covering challenges and advances in tools , resources , models , applications , and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18:1–52.

Ramy Baly, Mitra Mohtarami, James Glass, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. *arXiv*.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. volume 1 of *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261.

Pengyuan Chen, Kai Ye, and Xiaohui Cui. 2021. Integrating n-gram features into pre-trained model: A novel ensemble model for multi-target stance detection. volume 12893 LNCS, pages 269–279. Springer Science and Business Media Deutschland GmbH.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. volume 1 of *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Sardistance @ evalita2020: Overview of the task on stance detection in italian tweets. volume 2765 of *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–10.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv*.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 69–76.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies. ACL, pages 1163–1168.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. International Conference of the Cross-Language Evaluation Forum for European Languages, pages 75–87.

Santiago González-Carvajal and Eduardo C. Garrido-Merchán. 2020. Comparing bert against traditional machine learning text classification. *arXiv e-prints*.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Rumoureval 2019: Determining rumour veracity and support for rumours. Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 845–854.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. *arXiv*.

Andreas Hanselowski, Avinesh P.V.S., Benjamin Schiller, Felix Caspelherr, Debanjan * Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018).

Tomáš Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting stance in czech news commentaries. ITAT, pages 176–180.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics (ACL).

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. pages 92–104.

Jude Khouja. 2020. Stance prediction and claim verification: An arabic perspective. Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), pages 8–17.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 475–480.

Dilek Küçük and Fazli Can. 2018. Stance detection on tweets: An svm-based approach. *arXiv*, pages 1–13.

Dilek Küçük and C. A.N. Fazli. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. volume 10859 LNCS of *International Conference on Applications of Natural Language to Information Systems*, pages 15–27.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2020. #brexit: Leave or remain? the role of user's community and diachronic evolution on stance detection. *Journal of Intelligent and Fuzzy Systems*, 39:2341–2352.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021a. P-stance: A large dataset for stance detection in political domain. pages 2355–2365.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021b. Improving stance detection with multi-dataset learning and knowledge distillation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6332–6345.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. Proceedings of the World Wide Web Conference, WWW 2021, pages 3453–3464. Association for Computing Machinery, Inc.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17:1–23.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. volume 2 of *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017*, pages 551–557.

Parinaz Sobhani, Saif M Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (SEM 2016), pages 159–169.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018.

Mariona Taulé, M. Antónia Martín, Francisco Rangel, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. volume 1881 of *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, pages 157–177.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. 5th SwissText & 16th KONVENS Joint Conference 2020.

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*.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. *Natural language understanding and intelligent applications*, pages 907–916.

Elena Zotova, Rodrigo Agerri, and German Rigau. 2021. Semi-automatic generation of multilingual datasets for stance detection in twitter. *Expert Systems with Applications*, 170:1–29.