Detecting Stance of Authorities towards Rumors in Arabic Tweets: A Preliminary Study

No Author Given

No Institute Given

Abstract. A myriad of studies addressed the problem of rumor verification in Twitter by either utilizing evidence from the propagation networks or external evidence from the Web. However none of these studies exploited evidence from trusted authorities. In this paper, we define the task of detecting the stance of authorities towards rumors in tweets, i.e., whether a tweet from an authority agrees, disagrees, or is unrelated to the rumor. We believe the task is useful to augment the sources of evidence utilized by existing rumor verification systems. We construct and release the first Authority STance towards Rumors (AuSTR) dataset, where evidence is retrieved from authority timelines in Arabic Twitter. Due to the relatively limited size of our dataset, we study the usefulness of existing datasets for stance detection in our task. We show that existing datasets are quite useful for the task; however, they are clearly insufficient, which motivates the need to augment them with annotated data constituting stance of authorities from Twitter.

Keywords: Evidence · Claims · Social media

1 Introduction

Existing studies for rumor verification in social media exploited the propagation networks as a source of evidence, where they focused on the stance of replies [32, 22, 12, 33, 34, 8, 28], structure of replies [25, 26, 11, 13, 31, 18, 9], and profile features of retweeters [24]. Recently, Dougrez-Lewis et al. [16] proposed augmenting the propagation networks with evidence from the Web. To our knowledge, no previous research has investigated exploiting evidence for rumor verification in social media from trusted authority timelines. We believe that detecting stance of relevant authorities towards rumors can be a great asset to augment the sources of evidence utilized by existing rumor verification systems. It can also serve as a valuable tool for fact-checkers to automate their process of checking authority tweets to verify certain rumors.

In this paper, we conduct a preliminary study for detecting stance of authorities towards rumors spreading in Twitter in the Arab world. Exploiting sources of evidence for Arabic rumor verification in Twitter is still under-studied; existing studies exclusively focused on the tweet text for verification [17, 27, 2, 30, 5]. A notable exception is the work done by Haouari et al. [18] that utilized the replies, their structure, and repliers' profile features to verify Arabic COVID-19 rumors. Several studies addressed Arabic stance detection in Twitter; however, the target was a specific topic not rumors [14, 20, 6]. A few datasets for stance detection for Arabic claim verification were released recently, where the evidence is either news articles [10, 3] or manually-crafted sentences [21]. However, there is no dataset where the rumors are tweets and the evidence is retrieved from authority timelines, neither in Arabic nor in other languages. To fill this gap, the contribution of our work is four-fold: (1) we define the task of detecting stance of authorities towards rumors in tweets, (2) we construct and release the first Authority STance for Rumors (AuSTR) dataset,[1] (3) we present the first study on the usefulness of existing stance detection datasets for our task, and (4) we perform a failure analysis to gain insights for the future work on the task.

The remainder of this paper is organized as follows. We outline the construction methodology of AuSTR in Section 2. Our experimental setup is presented in Section 3. Finally, we discuss and analyze our results in Section 4.

2 Constructing AuSTR Dataset

To construct AuSTR where both the rumor and evidence are tweets, we exploit both fact-checking articles and variant authority Twitter accounts.

Exploiting Fact-checking Articles. Fact-checkers who attempt to verify rumors usually provide in their fact-checking articles some examples of social media posts (e.g., tweets) propagating the specific rumors, and other posts from trusted authorities that constitute evidence to support their verification decisions. To construct AuSTR, we exploit both examples of tweets: stating rumors and showing evidence from authorities as provided by those fact-checkers. Specifically, we used AraFacts [4], a large dataset of Arabic rumors collected from 5 fact-checking websites. From those rumors, we selected only the ones that are expressed in tweets and have evidence in tweets as well.[2] We then extracted the rumor-evidence pairs as follows. For *true* and *false* rumors, we selected a single tweet example and all provided evidence tweets, which are then labeled as having *agree* and *disagree* stances respectively.[3] If the fact-checkers provided the authority account but stated no evidence was found to support or deny the rumor, we selected one or two tweets from the authority timeline posted soon before the rumor time, and assigned the *unrelated* label to the pairs.

Exploiting Authority Accounts. Given that fact-checkers focus more on *false* rumors than *true* ones, we ended up with only 4 *agree* pairs as opposed to 118 *disagree* pairs following the above step. To further expand our *agree* pairs, we did the reverse of the previous approach, where we collected the evidence first. Specifically, we started from a set of Twitter accounts of authorities (e.g., ministers, presidents, embassies, organization accounts, etc.) covering most of

[1] Link for AuSTR is hidden for blind review.
[2] We contacted the authors of AraFacts to get this information as it was not released.
[3] We only kept evidence expressed in *text* rather than in image or video.

the Arab countries and multiple domains (e.g., politics, health, and sports), and selected recent tweets stating claims from their timelines. For each claim, we used Twitter search interface to look for tweets from regular users expressing it, but tried to avoid exact duplicates. Finally, to get closer to the real scenario, where percentage of *unrelated* tweets is usually higher than percentages of *agree* and *disagree* tweets in the authority timelines, we further expanded the *unrelated* pairs by selecting one or two *unrelated* recent tweets from the authority timeline posted before the rumor time for each *agree* and *disagree* pairs.

Overall, we end up with 409 pairs covering 171 unique claims, where 41 are *true* and 130 are *false*. Among those pairs, 118 are *disagree* (29%), 62 are *agree* (15%), and 229 are *unrelated* (56%).

3 Experimental Setup

Datasets. To study the usefulness of existing Arabic datasets that target stance for claim verification, we adopted the following ones for training:

1. **ANS [21]** of 3,786 **(claim, sentence)** pairs, where claims were extracted from news article titles from trusted sources, then annotators were asked to generate *true* and *false* sentences towards them by adopting paraphrasing and contradiction respectively. The sentences are annotated as either *agree*, *disagree*, or *other* towards the claims.
2. **ArabicFC [10]** of 3,042 **(claim, article)** pairs, where claims are extracted from a single fact-checking website verifying political claims about war in Syria, and articles collected by searching Google using the claim. The articles are annotated as either *agree*, *disagree*, *discuss*, or *unrelated* to the claim.
3. **AraStance [3]**: 3,676 **(claim, article)** pairs, where claims are extracted from 3 Arabic fact-checking websites covering multiple domains and Arab countries. The articles were collected and annotated similar to ArabicFC.

To train our models, we considered only three labels, namely, *agree*, *disagree*, or *unrelated*. For ANS and AraStance, we used the same data splits provided by the authors; however, we split the ArabicFC into 70%, 10%, and 20% of the claims for training, development, and testing respectively[4]. When splitting data, we assigned all pairs having the same claim to the same split. Table 1 shows the size of different data splits of the three datasets. Due to the limited size of AuSTR, in this work, we opt to utilize it only as a *test set* while using the above datasets for training to show their usefulness in our task.

Stance Models. To train our stance models, we fine-tuned BERT [15] to classify whether the evidence sentence/article *agrees* with, *disagrees* with, or is *unrelated* to the claim. We feed BERT the claim text as sentence A, the evidence as sentence B (truncated if needed) separated by the [SEP] token. Finally, we use

[4] We release ArabicFC splits for reproducibility.

Table 1. Data splits of the Arabic stance datasets used for training.

Label	ANS			ArabicFC			AraStance		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Agree	903	268	130	323	32	119	739	129	154
Disagree	1686	471	242	66	8	13	309	76	64
Unrelated	63	16	7	1464	198	410	1553	294	358
Total	2652	755	379	1853	238	542	2601	499	576

the contextual representation of the [CLS] token as input to a single classification layer with three output nodes, added on top of the BERT architecture to compute the probability for each class of stance.

Various Arabic BERT-based models were released recently [7, 29, 23, 19, 1]; we opted to choose ARBERT [1] as it was shown to achieve better performance on the stance datasets adopted in our work [3]. We adopted the authors' setup [3] by training the models for a maximum of 25 epochs, where early stopping was set to 5 and sequence length to 512. We trained 7 different models in an ablation study using different combinations of the stance datasets mentioned earlier.

4 Results and Discussion

The research question we address in this preliminary study is whether the existing stance detection datasets are useful or not in our task. To answer it, we use combinations of the existing datasets for training and AuSTR for testing. We also show how models trained on those combinations perform on their own corresponding in-domain test sets. While the results on the in-domain test sets are not comparable, since those test sets are different, they constitute an estimated upper bound performance. To evaluate the models, we report per-class F_1 and macro-F_1 scores. Table 2 presents the performance results of all experiments, which demonstrate several interesting observations.

First, we notice that almost all models (except a few) were able to achieve higher performance on their own in-domain test sets compared to AuSTR. This shows that domain adaptation was not very effective, and that AuSTR (and its corresponding task) might be more challenging.

Second, when using individual stance datasets for training, the model trained on AraStance clearly outperformed the others in all measures when tested on AuSTR. We note that ArabicFC is severely imbalanced, where the *disagree* class represents only 3.3% of the data, yielding a very poor performance on that class even when tested on its own in-domain test set. A similar conclusion was found by previous studies [10, 3]. As for ANS, evidence is manually crafted, which is not as realistic as tweets from authorities. Alternatively, AraStance claims are extracted from three fact-checking websites,[5] covering multiple domains and Arab countries, similar to AuSTR, and the evidence is represented in articles written by journalists, not manually crafted.

[5] Claims are collected from sources other than the ones we used to construct AuSTR.

Third, when tested on AuSTR, the model trained on all datasets combined exhibits the best performance on the *disagree* class; however its performance was severely degraded compared to the AraStance model on the *agree* class. This indeed needs further investigation.

Furthermore, we observe that AraStance achieved the highest F_1(D) when used solely for training, and whenever combined with the other datasets. To investigate this, we manually examined a 10% random sample of *disagreeing* training articles. We found they have common words such as *rumors*, *not true*, *denied*, and *fake*; similar keywords appear in some *disagreeing* tweets of AuSTR.

Finally, we observe that there is a clear discrepancy in the performance across different classes. Considering the model trained on all datasets for example, F_1(A) is 0.74 while F_1(D) is 0.65. Moreover, it is clear that detecting the *disagree* stance is the most challenging subtask, which we expect to benefit from in-domain training. Overall, we believe training and testing on tweets is very different, as they are very short and informal, which needs special pre-processing.

Table 2. Performance on both the in-domain test sets and AuSTR, measured in per-class F_1 (A: Agree, D: Disagree, U: Unrelated) and macro-F_1. On AuSTR, bold and underlined values indicate best and second-best performance respectively.

	Test on in-Domain Set				**Test on AuSTR**			
Training Set	F_1(A)	F_1(D)	F_1(U)	m-F_1	F_1(A)	F_1(D)	F_1(U)	m-F_1
ANS	0.824	0.901	0.923	0.882	0.653	0.578	0.709	0.647
ArabicFC	0.770	0.090	0.915	0.591	0.641	0.434	0.799	0.625
AraStance	0.898	0.833	0.95	0.894	**0.837**	0.613	<u>0.865</u>	**0.772**
ANS+ArabicFC	0.807	0.866	0.899	0.857	0.678	0.587	0.862	0.709
ANS+AraStance	0.893	0.909	0.955	0.919	0.743	0.629	0.847	0.740
ArabicFC+AraStance	0.765	0.555	0.897	0.739	<u>0.754</u>	<u>0.635</u>	0.862	0.750
All Three Datasets	0.778	0.742	0.889	0.803	0.741	**0.646**	**0.866**	<u>0.751</u>

Failure Analysis. We conducted a failure analysis on 17 examples from AuSTR that failed to be predicted correctly by *all* of our 7 trained models. We found that we can attribute the failures to two main reasons: (1) *Writing Style*, where the authority is denying a rumor about herself speaking in the first person. This constitutes 64.7% of the examined failures. We believe this is due to the fact that none of the stance datasets we used for training have evidence written by authorities themselves, as the source was either news articles written by journalists, or paraphrased or contradicted news headlines manually crafted by annotators. (2) *Indirect Disagreement/Agreement*, where the authority is indirectly denying/supporting the rumor. Examples of both types of failures are presented in Table 3. These findings motivate the need to augmenting existing stance datasets with rumor-evidence pairs from Twitter to further improve the performance of detecting the stance of authorities towards rumors from their tweets.

Table 3. Sample examples failed to be predicted correctly by **all** models. The golden label for the examples is either Agree or Disagree. Failure types are writing style, indirect disagreement, and indirect agreement for the examples in order.

Rumor tweet [posting date]	Evidence tweet [posting date]
Mortada Mansour passed away recently of a heart attack.[29-10-2021]	**@Mortada5Mansour**: I am having my dinner now, and after a few minutes I will share a voice and video to reassure you, and I will reply to those who disturbed my family members in my village and caused the anxiety to all my fans.[29-10-2021]
Egypt does not give a vaccine to its citizens, the Gulf countries sponsor them: Saudi Arabia / Sultanate of Oman / Qatar refuses their intervention, so there is no other than Kuwait, the country of humanity that receives them and feeds them. What is the mysterious secret? Kuwait treats Egypt with special treatment.[07-05-2021]	**@mohpegypt**: Information about the #coronavirus vaccine. To book a vaccine, please visit the website http://egcovac.mohp.gov.eg or go to the nearest health unit (for citizens who have difficulty registering online). For more information, please call the hotline: 15335 #together_rest_assured.[10-05-2021]
Urgent The headquarters of the fourth channel was stormed by the militias of the Sadrist movement in the capital, Baghdad.[04-11-2022]	**@MAKadhimi**: The attack on one of the Iraqi media outlets, and the threat to the lives of its employees, is a reprehensible act and represents the highest level of transgression against the law and freedom of the press and does not fall within the peaceful and legal practices and protests. We directed that the perpetrators be held accountable, and that protection be tightened on press institutions.[04-11-2022]

5 Conclusion and Future Work

In this paper, we defined the task of detecting stance of authorities towards rumors in tweets, and released the first dataset for the task targeting Arabic rumors. We studied the usefulness of existing Arabic datasets for stance detection for claim verification in our task. Based on our experiments and failure analysis, we found that although existing stance datasets showed to be quite useful for the task, they are obviously insufficient and there is a need to augment them with stance of authorities from Twitter data. In addition to expanding AuSTR to have sufficient training data for the task that can be use solely or to augment existing stance datasets, we plan to explore and contribute with stance models specific to the task.

References

1. Abdul-Mageed, M., Elmadany, A., et al.: Arbert & marbert: Deep bidirectional transformers for arabic. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 7088–7105 (2021)
2. Al-Yahya, M., Al-Khalifa, H., Al-Baity, H., AlSaeed, D., Essam, A.: Arabic fake news detection: comparative study of neural networks and transformer-based approaches. Complexity **2021**
3. Alhindi, T., Alabdulkarim, A., Alshehri, A., Abdul-Mageed, M., Nakov, P.: Arastance: A multi-country and multi-domain dataset of arabic stance detection for fact checking. NLP4IF 2021 p. 57 (2021)
4. Ali, Z.S., Mansour, W., Elsayed, T., Al-Ali, A.: Arafacts: the first large arabic dataset of naturally occurring claims. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 231–236 (2021)
5. Alqurashi, S., Hamoui, B., Alashaikh, A., Alhindi, A., Alanazi, E.: Eating garlic prevents covid-19 infection: Detecting misinformation on the arabic content of twitter. arXiv preprint arXiv:2101.05626 (2021)
6. Alqurashi, T.: Stance analysis of distance education in the kingdom of saudi arabia during the covid-19 pandemic using arabic twitter data. Sensors **22**(3), 1006 (2022)
7. Antoun, W., Baly, F., Hajj, H.: Arabert: Transformer-based model for arabic language understanding. In: LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020. p. 9 (2020)
8. Bai, N., Meng, F., Rui, X., Wang, Z.: A multi-task attention tree neural net for stance classification and rumor veracity detection. Applied Intelligence pp. 1–11 (2022)
9. Bai, N., Meng, F., Rui, X., Wang, Z.: Rumor detection based on a source-replies conversation tree convolutional neural net. Computing **104**(5), 1155–1171 (2022)
10. Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A., Nakov, P.: Integrating stance detection and fact checking in a unified corpus. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 21–27. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2004, https://aclanthology.org/N18-2004
11. Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., Huang, J.: Rumor detection on social media with bi-directional graph convolutional networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 549–556 (2020)
12. Chen, L., Wei, Z., Li, J., Zhou, B., Zhang, Q., Huang, X.J.: Modeling evolution of message interaction for rumor resolution. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6377–6387 (2020)
13. Choi, J., Ko, T., Choi, Y., Byun, H., Kim, C.k.: Dynamic graph convolutional networks with attention mechanism for rumor detection on social media. Plos one **16**(8), e0256039 (2021)
14. Darwish, K., Magdy, W., Zanouda, T.: Improved stance prediction in a user similarity feature space. In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017. pp. 145–148 (2017)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

16. Dougrez-Lewis, J., Kochkina, E., Arana-Catania, M., Liakata, M., He, Y.: Phemeplus: Enriching social media rumour verification with external evidence. In: Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER). pp. 49–58 (2022)
17. Elhadad, M.K., Li, K.F., Gebali, F.: Covid-19-fakes: A twitter (arabic/english) dataset for detecting misleading information on covid-19. In: International Conference on Intelligent Networking and Collaborative Systems. pp. 256–268. Springer (2020)
18. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.: Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 72–81 (2021)
19. Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., Habash, N.: The interplay of variant, size, and task type in arabic pre-trained language models. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 92–104 (2021)
20. Jaziriyan, M.M., Akbari, A., Karbasi, H.: Exaasc: A general target-based stance detection corpus in arabic language. In: 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE). pp. 424–429. IEEE (2021)
21. Khouja, J.: Stance prediction and claim verification: An Arabic perspective. In: Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER). Association for Computational Linguistics, Seattle, USA (2020)
22. Kumar, S., Carley, K.: Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy (Jul 2019)
23. Lan, W., Chen, Y., Xu, W., Ritter, A.: An empirical study of pre-trained transformers for Arabic information extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4727–4734. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.382, https://aclanthology.org/2020.emnlp-main.382
24. Liu, Y., Wu, Y.F.B.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
25. Ma, J., Gao, W., Wong, K.F.: Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics (2017)
26. Ma, J., Gao, W., Wong, K.F.: Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics (2018)
27. Mahlous, A.R., Al-Laith, A.: Fake news detection in arabic tweets during the covid-19 pandemic. International Journal of Advanced Computer Science and Applications **12**(6) (2021)
28. Roy, S., Bhanu, M., Saxena, S., Dandapat, S., Chandra, J.: gdart: Improving rumor verification in social media with discrete attention representations. Information Processing & Management **59**(3), 102927 (2022)
29. Safaya, A., Abdullatif, M., Yuret, D.: KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 2054–2059. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020). https://doi.org/10.18653/v1/2020.semeval-1.271, https://aclanthology.org/2020.semeval-1.271

30. Sawan, A., Thaher, T., Abu-el rub, N.: Sentiment analysis model for fake news identification in arabic tweets. In: 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT). pp. 1–6 (2021). https://doi.org/10.1109/AICT52784.2021.9620509
31. Song, C., Shu, K., Wu, B.: Temporally evolving graph neural network for fake news detection. Information Processing & Management **58**(6), 102712 (2021)
32. Wu, L., Rao, Y., Jin, H., Nazir, A., Sun, L.: Different absorption from the same sharing: Sifted multi-task learning for fake news 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). Association for Computational Linguistics, Hong Kong, China (Nov 2019)
33. Yu, J., Jiang, J., Khoo, L.M.S., Chieu, H.L., Xia, R.: Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1392–1401. Association for Computational Linguistics, Online (Nov 2020)
34. Yuan, C., Qian, W., Ma, Q., Zhou, W., Hu, S.: Srlf: A stance-aware reinforcement learning framework for content-based rumor detection on social media. arXiv preprint arXiv:2105.04098 (2021)