Social Network Analysis and Mining (2024) 14:34
https://doi.org/10.1007/s13278-023-01189-3

ORIGINAL ARTICLE

Are authorities denying or supporting? Detecting stance of authorities towards rumors in Twitter

Fatima Haouari[1] · Tamer Elsayed[1]

Received: 25 September 2023 / Revised: 25 September 2023 / Accepted: 20 December 2023

Abstract

Several studies examined the leverage of the stance in conversational threads or news articles as a signal for rumor verification. However, none of these studies leveraged the stance of *trusted authorities*. In this work, we define the task of detecting the stance of authorities towards rumors in Twitter, i.e., whether a tweet from an authority supports the rumor, denies it, or neither. We believe the task is useful to augment the sources of evidence exploited by existing rumor verification models. We construct and release the *first* Authority STance towards Rumors (AuSTR) dataset, where evidence is retrieved from authority timelines in Arabic Twitter. The collection comprises 811 (rumor tweet, authority tweet) pairs relevant to 292 unique rumors. Due to the relatively limited size of our dataset, we explore the adequacy of existing Arabic datasets of stance towards claims in training BERT-based models for our task, and the effect of augmenting AuSTR with those datasets. Our experiments show that, despite its limited size, a model trained solely on AuSTR with a class-balanced focus loss exhibits a comparable performance to the best studied combination of existing datasets augmented with AuSTR, achieving a performance of 0.84 macro-F1 and 0.78 F1 on debunking tweets. The results indicate that AuSTR can be sufficient for our task without the need for augmenting it with existing stance datasets. Finally, we conduct a thorough failure analysis to gain insights for the future directions on the task.

Keywords Claims · Evidence · Stance · Fact-checking · Social media

1 Introduction

Social media platforms (e.g., Twitter) have become a medium for rapidly spreading rumors along with emerging events Vosoughi et al. (2018). Those rumors may have a lasting effect on users' opinion even after it is debunked, and may continue influence them if not replaced with convincing evidence Nyhan and Reifler (2015). 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 Wu et al. (2019); Kumar and Carley (2019); Chen et al. (2020); Yu et al. (2020); Bai et al. (2022); Roy et al. (2022), structure of replies Ma et al. (2018); Bian et al. (2020); Choi et al. (2021); Song et al. (2021); Haouari et al. (2021); Bai et al. (2022), and profile features of retweeters Liu and Wu (2018). Recently, Dougrez-Lewis et al. (2022) proposed augmenting the propagation networks with evidence from the Web, and Hu et al. (2023) proposed exploiting both text and images retrieved from the Web as sources of evidence. A large body of existing studies in the broader literature have examined exploiting the stance of conversational threads Zubiaga et al. (2016); Derczynski et al. (2017) or news articles Ferreira and Vlachos (2016); Alhindi et al. (2021) towards claims as a signal for verification.

However, to our knowledge, no previous research has investigated exploiting evidence from the timelines of *trusted authorities* for rumor verification in social media.

This article presents a major extension of a previous work published at ECIR 2023 Haouari and Elsayed (2023). Extensions include (1) expanding the dataset by doubling the number of examples, (2) proposing a new semi-automated approach for collecting the data, (3) studying the usefulness of two more Arabic stance datasets, (3) using in-domain data for training the models, (4) fine-tuning our BERT models over different hyper-parameters, and (5) investigating various loss functions to alleviate the class-imbalance issue.

✉ Fatima Haouari
200159617@qu.edu.qa

Tamer Elsayed
telsayed@qu.edu.qa

[1] Computer Science and Engineering Department, Qatar University, Doha, Qatar

Published online: 26 January 2024

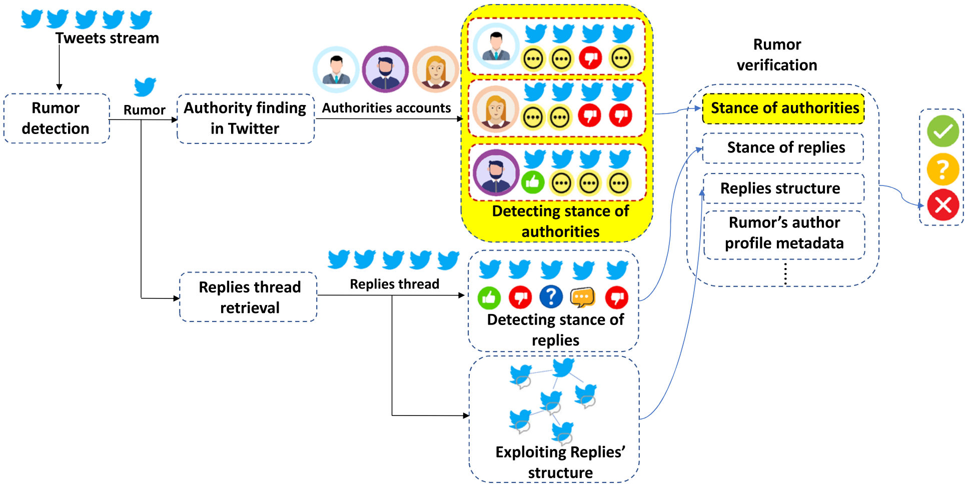


Fig. 1 Positioning the stance of authorities detection task (highlighted in yellow) in the rumor verification pipeline (color figure online)

An authority is *an entity with the real knowledge or power to verify or deny a specific rumor* Haouari et al. (2023); Haouari and Elsayed (2023). Therefore, 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 verifying rumors from authorities.

In this work, we address the problem of *detecting the stance of authorities towards rumors in Twitter*, defined as follows: Given a rumor expressed in a tweet and a tweet posted by an authority of that rumor, detect whether the tweet *supports* (agrees with) the rumor, *denies* (disagrees with) it, or *not* (other). Figure 1 presents our perception of the role of detecting the stance of authorities in a typical pipeline of rumor verification over Twitter. Given a rumor expressed in a tweet, both the reply thread and the corresponding authority Twitter accounts are retrieved. The replies structure, the replies stance, and the authorities stance in addition to other potential signals will then be exploited by the rumor verification model to decide the veracity of the rumor. In our work, we assume that the authorities for a given rumor are already retrieved Haouari et al. (2023), and we only target the detection of the stance of those authorities towards the rumors. In particular, our model is supposed to do so over the tweet timelines of the corresponding retrieved authorities. While being very important source of evidence for rumor verification, it is worth mentioning that stance of authorities can complement other sources, especially if authorities are automatically retrieved, thus not fully accurate.

A closer look at the literature on *Arabic* rumor verification in Twitter in particular reveals that utilizing signals for verification is under-explored; most existing studies relied on the tweet textual content to detect its veracity Hasanain et al. (2020); Elhadad et al. (2020); Mahlous and Al-Laith (2021); Al-Yahya et al. (2021); Alqurashi et al. (2021); Sawan et al. (2021). Some notable exceptions are the work done by Albalawi et al. (2023) (who exploited the images and videos embedded in the tweet), the study done by Haouari et al. (2021) (who used the reply thread structure and reply network signals), and the work done by Althabiti et al. (2022) (who proposed detecting sarcasm and hate speech in the replies for Arabic rumor verification in Twitter).

To fill this literature gap, we first introduce the problem of detecting the stance of authorities towards rumors in Twitter. We then construct the first dataset for the task and release it along with its construction guidelines to facilitate future research. Moreover, we investigate the usefulness of existing Arabic stance datasets towards claims for our task. Finally, we explore the mitigation of the traditional class-imbalance issue in stance datasets by experimenting with various loss functions. Our experiments show that training a model with our dataset solely, despite being relatively very small, exhibits a performance that is (at least) on bar with training with other (combinations of) existing stance datasets, indicating that existing stance datasets are not really needed for the task. The contributions of this paper are as follows:

1. We introduce and define the task of detecting the stance of authorities towards rumors that are propagating in Twitter.
2. We release the first Authority STance towards Rumors (AuSTR) dataset for that specific task[1] targeting the *Arabic* language.
3. We explore the adequacy of existing Arabic datasets of stance towards claims for our task, and the effect of augmenting our in-domain data with those datasets on the performance of the model.
4. We investigate the performance of the models when adopting variant loss functions to alleviate the class-

[1] https://github.com/Fatima-Haouari/AuSTR

imbalance issue, and we perform a thorough failure analysis to gain insights for the future work on the task.

The rest of this paper is organized as follows. We present our literature review in Sect. 2 and define the problem. We are targeting in this work in Sect. 3. In Sect. 4, we present our dataset construction approach. Our experimental approach is presented in Sect. 5. We discuss the experimental setup in Sect. 6 and thoroughly analyze the results and answer the research questions in Sect. 7. We conduct a failure analysis to gain insights for future directions and discuss the limitations of our study in Sect. 8. Finally, we conclude and suggest some future directions in Sect. 9.

2 Related work

In this section, we briefly review the related studies to our work. Specifically, we review rumor debunking in social media studies in Sect. 2.1, we give an overview of studies for stance detection for claim verification in Sect. 2.2, and we review authorities for rumor verification studies in Section. 2.3.

2.1 Rumor debunking in social media

Several studies on rumors debunking in Twitter suggested exploiting online debunkers, i.e., users who share fact-checking URLs to stop the propagation of a circulating rumor Vo and Lee (2018, 2019, 2020a, 2020b); You et al. (2019); Mu et al. (2022). To encourage online debunkers in Twitter remain engaged in correcting rumors, some studies proposed fact-checking URLs recommender systems Vo and Lee (2018); You et al. (2019). Vo and Lee (2019, 2020b) proposed a fact-checking response generator framework to stop the propagation of fake news, and exploited the replies of users who usually debunk rumors in Twitter to implement their model. Vo and Lee (2020a) on the other hand introduced a multimodal framework to retrieve fact-checking articles to be incorporated into rumor spreaders conversations threads to discourage propagating rumors in social media.

Differently, in our work we consider authorities as credible debunkers who may post tweets supporting or debunking a specific rumor circulating in Twitter.

2.2 Stance detection for claim verification

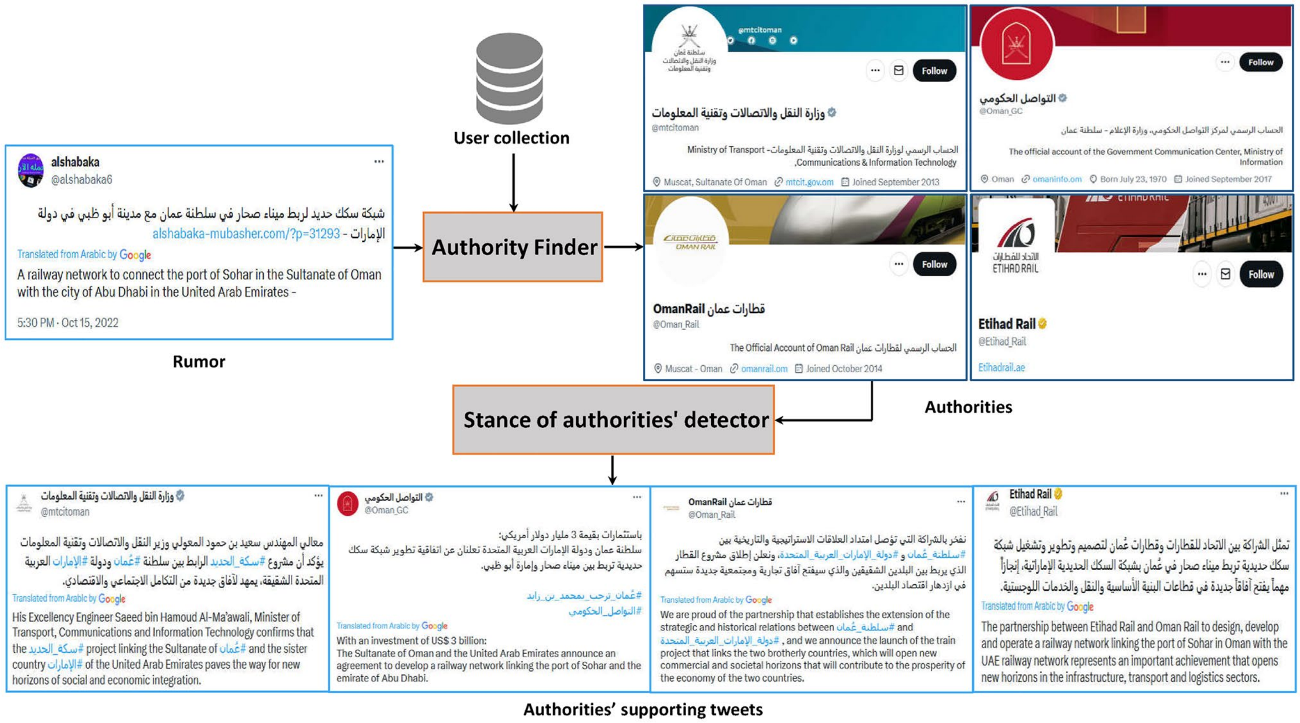
A myriad of studies have investigated detecting the stance towards claims to identify its veracity Hardalov et al. (2022). Some focusing on detecting the stance of conversation threads in social media Zubiaga et al. (2016); Derczynski et al. (2017); Gorrell et al. (2019), and others on the stance of news articles Ferreira and Vlachos (2016); Pomerleau and Rao (2017); Baly et al. (2018); Alhindi et al. (2021). Existing studies either considered the stance as an isolated module in the verification system Zubiaga et al. (2016); Ferreira and Vlachos (2016); Derczynski et al. (2017); Gorrell et al. (2019), or considered the stance of the evidence towards the claim as the veracity label Thorne et al. (2018); Hanselowski et al. (2018); Guderlei and Aßenmacher (2020); Slovikovskaya and Attardi (2020). Multiple approaches were proposed recently considering verification as stance detection, mainly targeting stance of articles towards claims, by either exploiting transformer-based models Slovikovskaya and Attardi (2020); Khouja (2020); Alhindi et al. (2021), or graph neural networks Zhou et al. (2019); Liu et al. (2020); Si et al. (2021). In the other hand, studies considering stance detection as a standalone component in the verification pipeline are mainly targeting the stance of conversation threads towards rumors in social media. A plethora of models were proposed to detect the stance of conversation threads such as tree and hierarchical transformers proposed by Ma and Gao (2020) and Yu et al. (2020), respectively.

A few studies addressed stance detection for Arabic claim verification recently, where the evidence is either news articles Baly et al. (2018); Alhindi et al. (2021) or manually crafted sentences from articles headlines Khouja (2020). In contrast, in our work, we define the task of detecting the authorities stance towards Arabic rumors where we consider it as a standalone component in the rumor verification pipeline, and we release the first dataset for the task. We study the usefulness of existing Arabic stance towards claims datasets for the task, and we evaluate the performance of the stance models when incorporating in-domain data for training the models. Finally, we investigate two loss functions who showed promising results to alleviate the class-imbalance issue identified as a major challenge for stance detection for rumor verification Li and Scarton (2020).

2.3 Authorities for rumor verification

A closer look to the literature on rumor verification in social media reveals that no study to date has examined exploiting evidence from authorities. Existing studies for rumor verification in social media exploited evidence from the propagation networks Liu and Wu (2018); Bai et al. (2022); Roy et al. (2022); Song et al. (2021); Haouari et al. (2021), Web Dougrez-Lewis et al. (2022), and stance of conversational threads Zubiaga et al. (2016); Derczynski et al. (2017); Gorrell et al. (2019).

Recently, Haouari et al. (2023) introduced the task of authority finder in Twitter which they define as follows: *given a tweet stating a rumor, retrieve a ranked list of authority accounts from Twitter that can help verify the rumor, i.e., they may tweet evidence that supports or denies the rumor*. The authors released the first Arabic test


User collection
Authority Finder
Rumor
Authorities
Stance of authorities' detector
Authorities' supporting tweets

alshabaka
@alshabaka6

شبكة سكك حديد لربط ميناء صحار في سلطنة عمان مع مدينة أبو ظبي في دولة
الإمارات - alshabaka-mubasher.com/?p=31293

Translated from Arabic by Google

A railway network to connect the port of Sohar in the Sultanate of Oman with the city of Abu Dhabi in the United Arab Emirates -

5:30 PM · Oct 15, 2022

وزارة النقل والاتصالات وتقنية المعلومات
@mtcitoman

الحساب الرسمي لوزارة النقل والاتصالات وتقنية المعلومات- Ministry of Transport ,Communications & Information Technology

Muscat, Sultanate Of Oman · mtcit.gov.om · Joined September 2013

التواصل الحكومي
@Oman_GC

الحساب الرسمي لمركز التواصل الحكومي، وزارة الإعلام - سلطنة عمان

The official account of the Government Communication Center, Ministry of Information

Oman · omaninfo.om · Born July 23, 1970 · Joined September 2017

OmanRail قطارات عمان
@Oman_Rail

الحساب الرسمي لقطارات عمان The Official Account of Oman Rail

Muscat - Oman · omanrail.om · Joined October 2014

Etihad Rail
@Etihad_Rail

Etihadrail.ae

وزارة النقل والاتصالات وتقنية المعلومات
@mtcitoman

معالي المهندس سعيد بن حمود المعولي وزير النقل والاتصالات وتقنية المعلومات يؤكد أن مشروع #سكة_الحديد الرابط بين سلطنة #عُمان ودولة #الإمارات العربية المتحدة الشقيقة، يمهد لآفاق جديدة من التكامل الاجتماعي والاقتصادي.

Translated from Arabic by Google

His Excellency Engineer Saeed bin Hamoud Al-Ma'awali, Minister of Transport, Communications and Information Technology confirms that the #سكة_الحديد project linking the Sultanate of #عُمان and the sister country #الإمارات of the United Arab Emirates paves the way for new horizons of social and economic integration.

التواصل الحكومي
@Oman_GC

باستثمارات بقيمة 3 مليار دولار أمريكي:
سلطنة عمان ودولة الإمارات العربية المتحدة تعلنان عن اتفاقية تطوير شبكة سكك حديدية تربط بين ميناء صحار وإمارة أبو ظبي.

#عُمان_ترحب_بمحمد_بن_زايد
#التواصل_الحكومي

Translated from Arabic by Google

With an investment of US$ 3 billion:
The Sultanate of Oman and the United Arab Emirates announce an agreement to develop a railway network linking the port of Sohar and the emirate of Abu Dhabi.

OmanRail قطارات عمان
@Oman_Rail

نفخر بالشراكة التي تؤصل امتداد العلاقات الاستراتيجية والتاريخية بين #سلطنة_عُمان و #دولة_الإمارات_العربية_المتحدة، ونعلن إطلاق مشروع القطار الذي يربط بين البلدين الشقيقين والذي سيفتح آفاق تجارية ومجتمعية جديدة ستسهم في ازدهار اقتصاد البلدين.

Translated from Arabic by Google

We are proud of the partnership that establishes the extension of the strategic and historical relations between #سلطنة_عُمان and #دولة_الإمارات_العربية_المتحدة , and we announce the launch of the train project that links the two brotherly countries, which will open new commercial and societal horizons that will contribute to the prosperity of the economy of the two countries.

Etihad Rail
@Etihad_Rail

تمثل الشراكة بين الاتحاد للقطارات وقطارات عُمان لتصميم وتطوير وتشغيل شبكة سكك حديدية تربط ميناء صحار في عُمان بشبكة السكك الحديدية الإماراتية، إنجازاً مهماً يفتح آفاقاً جديدة في قطاعات البنية الأساسية والنقل والخدمات اللوجستية.

Translated from Arabic by Google

The partnership between Etihad Rail and Oman Rail to design, develop and operate a railway network linking the port of Sohar in Oman with the UAE railway network represents an important achievement that opens new horizons in the infrastructure, transport and logistics sectors.

Fig. 2 An example of a rumor along with its corresponding authorities and a set of *supporting* tweets detected from the authorities timelines (The example is from our constructed AuSTR dataset)

collection for the task and proposed a hybrid model that exploits both lexical, semantic, and user networks signals to find authorities. The authority finder task was then introduced as part of the CheckThat! 2023 lab shared tasks Barrón-Cedeño et al. (2023); Haouari et al. (2023), and it was deployed as a system component as part of a live system for Arabic claim verification Sheikh Ali et al. (2023). Differently, in our work we assume that the authority is already retrieved, and the task is to detect the stance of her tweets towards a given rumor.

3 Overview of our work

Figure 2 shows an example of a rumor about an establishment of a new railway to connect the Sultanate of Oman and the United Arab of Emirates (UAE). We assume that the authorities for this rumor are retrieved by an "authority finding" model (here some of the highly relevant authorities are the ministry of transport in Oman, the Omani government communication center, and both Oman's and UAE's rails projects). The figure shows an example tweet from each of the timelines of the authorities that actually supports the rumor.[2]

In this work, we introduce the task of detecting the stance of authorities towards rumors in Twitter. Due to the lack of datasets for the task, we construct and release the first Authority STance towards Rumors (AuSTR) dataset (Sect. 4). We exploit both fact-checking articles and authority Twitter accounts to *manually* collect *debunking*, *supporting*, and *other* (rumor tweet, authority tweet) pairs. Additionally, we propose a semi-automated approach utilizing the Twitter search API to further expand our *debunking* pairs.

Due to the limited size of our dataset, we investigate the usefulness of existing datasets of stance towards Arabic claims (Sect. 7.1 and Sect. 7.2). Adopting a BERT-based stance model, we perform extensive experiments using five variant Arabic stance datasets, where the target is a claim but the context is either an article, article headline, or a tweet, to investigate if the stance model trained with each of them is able to generalize to our task. We then explore the effect of augmenting our in-domain data with each of the Arabic stance datasets on the performance of the model (Sect. 7.3). To mitigate the class-imbalance issue, we explore variant

[2] This is an example from AuSTR that actually has 11 supporting tweets overall.

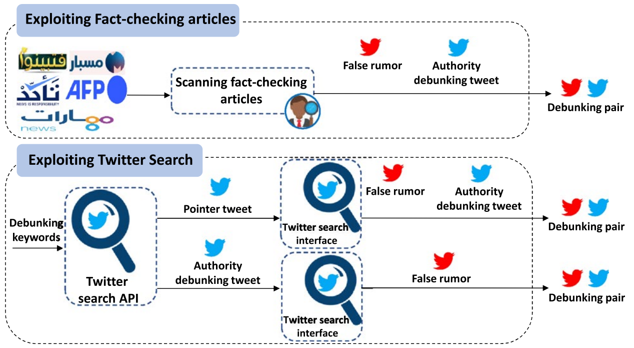


Fig. 3 Our approach for collecting AuSTR *debunking* pairs

loss functions replacing the cross-entropy loss (Sect. 7.4). Finally, we conduct a thorough error analysis to gain insights for the future improvements (Sect. 8.1).

4 Constructing AuSTR dataset

To address the lack of datasets of authority stance towards rumors, in this work, we introduce the *first* Authority STance towards Rumors (denoted as *AuSTR*) dataset. Our focus is on Arabic, as it is one of the most popular languages in Twitter Alshaabi et al. (2020), yet it is under-explored for rumor verification. Our dataset consists of 811 pairs of rumors (expressed in tweets) and authority tweets related to 292 unique rumors. Tweets of authorities are labeled as either *disagree*, *agree*, or *other*, as defined earlier. To construct AuSTR, we collected the *debunking* pairs manually (details in Sect. 4.1) by exploiting fact-checking articles and adopting a semi-automated approach. *Supporting* pairs were collected by manually exploring authority accounts and the Twitter search interface, in addition to utilizing the fact-checking articles (details in Sect. 4.2). Finally, to collect our *other* pairs we manually examined the timelines of the authorities of our *debunking* and *supporting* pairs to select tweets that are neither agreeing nor disagreeing with the rumor, in addition to exploiting fact-checking articles (details in Sect. 4.3).

4.1 Collecting debunking pairs

Figure 3 depicts an overview of our approach to construct the *debunking* pairs of AuSTR. We leveraged both the fact-checking articles and a semi-automated approach which we propose in this work.

4.1.1 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, along with other posts from trusted authorities that constitute evidence to support their verification decisions. For AuSTR, we exploit both examples of tweets: stating rumors and showing evidence from authorities as provided by those fact-checkers. Specifically, we used AraFacts Ali et al. (2021), a large dataset of Arabic rumors collected from five fact-checking websites. From those rumors, we selected only the ones that are expressed in tweets and for which the fact-checkers provided evidence in tweets as well.[3] For *false* rumors, we selected a single tweet example of the rumor and all provided evidence tweets for it, which are then labeled as having *disagree* stances. Adopting this approach, we ended up with 118 *debunking* pairs.

4.1.2 Exploiting Twitter search

Additionally, we adopted a semi-automated approach to collect more debunking pairs using Twitter search. First, we used the Twitter Academic API[4] to collect *potentially debunking* tweets, i.e., tweets with denying keywords and phrases such as "*fake news*", "*fabricated*", "*rumors*", and "*denied the news*". Specifically, we used 21 keywords/phrases[5] to search Twitter to retrieve Arabic tweets from the period of July 1, 2022, to December 31, 2022. To narrow down our search and reduce the noisy tweets, we excluded retweets and the tweets of non-verified accounts. Given that fact-checkers usually use most of these keywords to debunk rumors, we also excluded tweets from verified Arabic fact-checking Twitter accounts. By adopting this approach, we were able to collect either debunking tweets from authorities themselves, or just *pointer* tweets from journalists or news agencies. For both types, we retrieved the rumor tweets by searching Twitter user interface using the main keywords in the debunked rumor by the authorities. For the later type, we manually examined the timelines of authorities to get the debunking tweets.

Table 1 presents examples of *debunking* tweets from authorities along with the search keywords used to retrieve them. An example of automatically retrieved pointer tweet and the manually collected *disagree* pair is presented in Table 2.

[3] We contacted the authors of AraFacts to get this information as it was not released.

[4] https://developer.twitter.com/en/products/twitter-api/academic-research.

[5] We release the keywords we used for collecting the debunking tweets in our data repository.

Table 1 Examples of *debunking* authority tweets (and their English translations) collected using the semi-automated approach along with the search keywords

Search keywords	Example of a collected tweet
غير صحيح	بيان من #غد الثورة: نشرت احد المواقع خبر غير صحيح عن قرار الحزب بالدعوة لحراك ١١ ١١ ...
Incorrect	**@AymanNour:** Statement from #Ghad El Thawra: One of the sites published incorrect news about the party's decision to call for the 11/11 movement ...
خبر كاذب	نفي خبر كاذب نشرته إحدى الصحف اللبنانية حول توقيف شقيق اللواء عثمان
Fake news	**@LebISF:** Denying a fake news published by a Lebanese newspaper about the arrest of Major General Othman's brother
عارٍ عن الصحة	.... أنباء عن (اختفاء مواطن أمريكي في وسط أو جنوب العراق بظروف غامضة، يعمل صحفيّاً) نؤكد أن هذا الخبر عارٍ عن الصحة...
Untrue	**@IraqiSpoxMOD:** ... news about (the disappearance of an American citizen in central or southern Iraq, under mysterious circumstances, who works as a journalist). We confirm that this news is untrue ...
مفبرك	...خطاب اعتراض الأهلي على زي الزمالك في السوبر مفبرك...
Fabricated	**@AlAhlyTV:** ...Al-Ahly's objection speech about Zamalek club uniforms in the super is fabricated...
شائعات	#بيان: تسري شائعات مفادها أن المديرية العامة للأمن العام أوقفت المواطنة سالي حافظ التي اقتحمت أحد المصارف في بيروت...
Rumors	**@DGSGLB**: #Statement: rumors are circulating that the General Directorate of General Security arrested Sally Hafez, who broke into a bank in Beirut...

Table 2 An example of an automatically collected *pointer debunking* tweet along with its manually collected *debunking* pair (with their English translation)

Tweet type	Tweet text
Pointer	السفارة القطرية في تونس: غير صحيح.. مقتل قطري بالمدينة العتيقة في بنزرت @naharkw: The Qatari Embassy in Tunisia: Incorrect.. A Qatari was killed in the ancient city of Bizerte. [11-08-2022]
Authority	تنفي سفارة دولة قطر لدى الجمهورية التونسية، ما تداولته وسائل إعلام عن أن المجنى عليه في حادثة بنزرت يحمل الجنسية القطرية، وتعبر عن تعازيها لأسرة الضحية وذوية. @QatarEmb_Tunis: The Embassy of the State of Qatar in the Republic of Tunisia denies what was reported by the media that the victim in the Bizerte incident holds Qatari nationality, and expresses its condolences to the victim's family and relatives. [11-08-2022]
Rumor	مقتل قطري في تونس يهز المدينة العتيقة بـ بنزرت #تونس @USER: The killing of a Qatari in Tunisia shakes the ancient city of Bizerte #Tunisia [12-08-2022]

4.2 Collecting supporting pairs

To collect *supporting* pairs, we adopted two approaches as presented in Fig. 4. Given that fact-checkers focus more on *false* rumors than *true* ones, exploiting fact-checking articles was not sufficient to collect *supporting* tweets, as adopting this approach, we were able to collect only 4 *agree* pairs as opposed to 118 *disagree* pairs. Thus, we manually collected a set of *governmental Arabic Twitter accounts* representing authorities related to health and politics, such as ministries and ministers, embassy accounts, and Arabic sports organizations accounts (e.g., football associations and clubs). Starting from 172 authority accounts from multiple Arabic countries,[6] we manually checked the timelines of those authorities from the period of July 1, 2022, to December 31, 2022. We selected *check-worthy* tweets, i.e, tweets containing verifiable claims that we think will be of general interest Shaar et al. (2021), and consider them as authority *supporting* tweets. We then used the main keywords in each claim to search Twitter through the user interface and selected a tweet propagating the same claim while avoiding near-duplicates. We ended up with 148 *agree* pairs in total. Table 3 shows

[6] We release our collected authority Twitter accounts in our data repository.

Table 3 An example of manually collected *supporting* authority tweet and a relevant rumor tweet expressing the same claim

Tweet type	Tweet text
Authority	الإعلام الأمني: اسعاف مقيم حاول الانتحار عن طريق طعن نفسه داخل احدى المساجد، وتم التحفظ على ألشخص وجاري اتخاذ الإجراءات القانونية اللازمة بالواقعة. @Moi_kuw: A resident who tried to commit suicide by stabbing himself inside a mosque was first aided, and the person was kept and the necessary legal measures are being taken in the incident. [04-12-2022]
Rumor	متداول #محاولة _انتحار : حاول الانتحار داخل مسجد الغانم بقرطبة ومازالت الاسباب مجهولة. @USER: Circulating #suicide_attempt: He attempted suicide inside Al-Ghanim Mosque in Cordoba, and the reasons are still unknown.[04-12-2022]

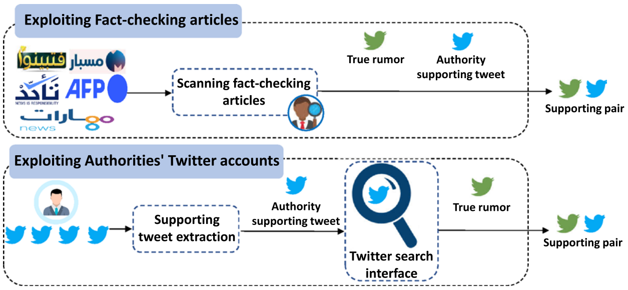

Exploiting Fact-checking articles
فتبينوا
مسبار
تأكد
NEWS IS RESPONSIBILITY
AFP
مهارات news
Scanning fact-checking articles
True rumor
Authority supporting tweet
Supporting pair
Exploiting Authorities' Twitter accounts
Supporting tweet extraction
Authority supporting tweet
Twitter search interface
True rumor
Supporting pair

Fig. 4 Collecting AuSTR *supporting* pairs approach

an example of a *supporting* authority tweet along with a relevant rumor.

4.3 Collecting other pairs

For some rumors, fact-checkers provide the authority account in their fact-checking article, but they state that no evidence was found to support or deny the rumor. For this case, we selected one or two tweets from the authority timeline posted soon before the rumor time, and assigned the *other* label to those pairs.

In reality, most of the tweets in authority timelines are neither supporting nor denying a given rumor. To get closer to that real scenario, for each *agree* and *disagree* pair, we manually examined the timeline of the authority within the same time period of the rumor, and selected at most two tweets, where we give higher priority to tweets related to the rumor's topic or at least have an overlap in some keywords with the rumor. A tweet of those is then labeled as *other* if it is either relevant to the rumor but is neither disagreeing nor agreeing with it, or it is completely irrelevant to it. We ended up with 466 *other* pairs.

It is worth noting that the evidence from authorities is not always expressed in the textual body of the tweet. We considered the case when some authorities may post evidence as an announcement embedded in an image or video.

Table 4 AuSTR statistics

Class	Pairs
Disagree	197 (24.3%)
Exploiting fact-checking articles	118
Semi-automated approach	79
Agree	148 (18.2%)
Exploiting fact-checking articles	4
Exploiting authorities accounts	144
Other	466 (57.5%)
Exploiting fact-checking articles	158
Manual	308
Total	811

4.4 Data quality

We present our dataset statistics in Table 4. Our data were annotated by one of the authors, a PhD candidate and native Arabic speaker working on rumor verification in Twitter. To measure the quality of our data, we randomly picked 10% of the pairs and asked a *second* annotator, a PhD holder and native Arabic speaker, to label them. The computed Cohen's Kappa for inter-annotator agreement Cohen (1960) was found to be 0.86, which indicates "almost perfect" agreement Landis and Koch (1977).

5 Experimental design

Due to the limited size of AuSTR, one of the main objectives of this work is to study the adequacy of using *existing* datasets of stance towards claims in training models for our task. Specifically, the goal is to first study whether models trained with existing stance datasets perform well on detecting the stance of authorities in particular, then investigate whether augmenting them with AuSTR improve the performance of those models. Moreover, since a major challenge of stance classification is the class-imbalance problem in the data Li and Scarton (2020), we also aim to explore whether incorporating different loss functions can mitigate that issue to further improve the performance of the models.

Accordingly, we aim to answer the following research questions:

- **RQ1:** To what extent will stance models trained with existing stance datasets be able to generalize to the task of detecting the stance of *authorities*?
- **RQ2:** What is the effect of combining all existing stance datasets for training?
- **RQ3:** Will training a stance model with AuSTR solely be sufficient? Will augmenting AuSTR with existing stance datasets for training improve the performance?
- **RQ4:** Will adopting different loss functions mitigate the class-imbalance problem thus improve the performance?

To address those research questions, we design our experiments as follows:

- **Cross-domain experiments** denote the case where existing datasets of stance towards claims are exploited for training. Each of the stance datasets is first used solely for training our models, then all datasets were aggregated and used for training. We refer to the datasets of stance towards claims as *cross-domain* datasets in the rest of the paper.
- **In-domain experiments** denote the case where AuSTR is used solely for training. We refer to AuSTR as *in-domain* dataset.
- **In-domain-augmented experiments** denote the case where AuSTR is augmented with existing datasets of stance towards claims. In those experiments, we study the effect of augmenting AuSTR with each of the cross-domain datasets separately, in addition to augmenting it with all of them.
- **Class-Imbalance experiments** denote the case where we adopt different loss functions that showed promising results earlier in the literature, to alleviate the class-imbalance problem.

6 Experimental setup

In this section, we present the setup we adopted to conduct our experiments.

Datasets

To study the adequacy of existing Arabic datasets of stance detection towards claims for the task of detecting the stance of authorities, we adopted the following five existing datasets in training:

- **ArCOV19-Rumors** Haouari et al. (2021) consists of 9,413 **tweets** relevant to 138 COVID-19 Arabic **rumors** collected from 2 Arabic fact-checking websites.

  We considered the tweets *expressing the rumor* as supporting (agree), the ones that are *negating the rumor* as denying (disagree), and the ones discussing the rumor but neither expressing nor negating it as *other*.
- **STANCEOSAURUS** Zheng et al. (2022) consists of 4,009 **(rumor, tweet)** pairs. The data cover 22 Arabic rumors collected from 3 Arabic fact-checking websites along with tweets, collected by the authors that are relevant to the rumors. The relevant tweets were annotated by their stance towards the rumor as either *supporting* (agree), *refuting* (disagree), *discussing*, *querying*, or *irrelevant*. In our work, we considered the last three labels as *other*.
- **ANS** Khouja (2020) consists of 3,786 **(claim, manipulated claim)** pairs, where claims were extracted from news article headlines 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*.
- **ArabicFC** Baly et al. (2018) consists of 3,042 **(claim, article)** pairs, where claims are extracted from a single fact-checking website verifying political claims about the 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. In our work, we considered the last two labels as *other*.
- **AraStance** Alhindi et al. (2021) consists of 4,063 **(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.

Figure 5 presents the per-class statistics for each dataset (including AuSTR), and Table 5 shows an example of a debunking text from each of them.

Data splits

Given that AuSTR constitutes only 811 pairs, we adopt cross-validation for evaluating our models. We randomly split it into five folds while assigning all pairs that are relevant to the same rumor to the same fold to avoid label leakage across folds.

For all of our models, whether AuSTR is exploited for training or not, we both *tune* and *test* only on folds from AuSTR; a single AuSTR fold (dev fold) is used for tuning the models and another (test fold) was used for testing. If AuSTR is used for training, the remaining three folds (training folds) are used for that purpose. When the cross-domain datasets are used for training, they are fully used for that purpose (and none of them is used for tuning nor testing).

Table 5 *Debunking* examples (and their English translations) from the cross-domain datasets

Dataset	Tweet text
ArCOV19-Rumors	لا صحة لما يتم تداوله عن تعرض لاعب يوفنتوس باولو ديبالا لعدوى فيروس كورونا، و مصدر الشائعة قناة فنزويلية. @USER: There is no truth to what is being circulated about Juventus player Paulo Dybala being infected with the coronavirus, and the source of the rumor is a Venezuelan channel. [13-03-2020]
STANCEOSAURUS	كنت ابكي على موت كاظم الساهر وطلع اخوه اللي مات. @USER: I was crying over the death of Kadim Al Sahir and it turned out that who died is his brother. [13-01-2022]
ANS	أصدر القضاء المغربى حكما بسجن الزفزافى لمدة ٢٠ عاما. The Moroccan judiciary issued a 20-year prison sentence for Zefzafi
ArabicFC	هيئة تحرير الشام تنفي إصابة قائدها الجولاني في ضربة روسية الجزيرة مباشر الأربعاء ٤ أكتوبر ٢٠١٧ … Hayat Tahrir al-Sham denies that its commander al-Julani was injured in a Russian strike, Al-Jazeera Mubasher, Wednesday, October 4, 2017..
AraStance	الفيديو المتداول بعنوان انفجار جوال في جيب أحد الأشخاص في أحد مراكز دبي التجارية لا صحة له بل حدث قبل أيام في مدينة أغادير في المغرب… The circulating video entitled “a mobile phone explosion in a person’s pocket in a Dubai mall” is not true. Rather, it happened a few days ago in the city of Agadir in Morocco..

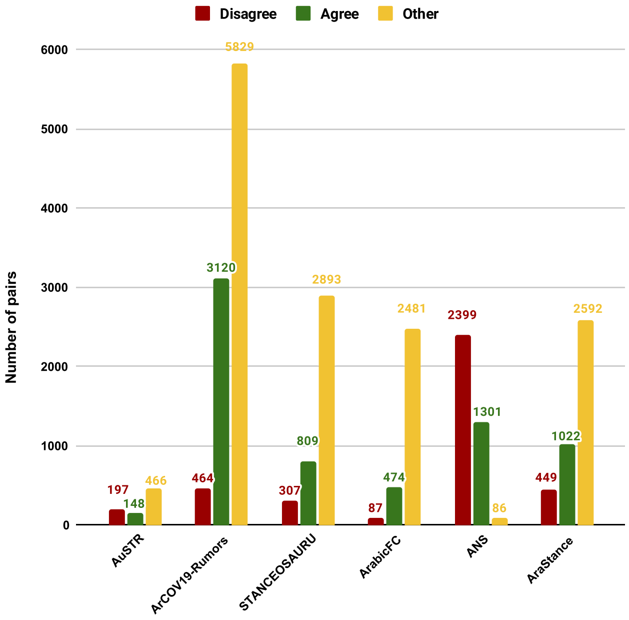


Fig. 5 Per-class statistics of cross-domain datasets adopted in our work, as well as AuSTR for comparison

For each experiment, we train five models to test on the five different folds of AuSTR, and finally report the average performance of the five models.

Stance models

To train our stance models, we fine-tuned BERT Devlin et al. (2018), following recent studies that adopted transformer-based models for stance detection Alhindi et al. (2021); Alturayeif et al. (2022); Zheng et al. (2022); Ng and Carley (2022) to classify whether the evidence *agrees* with the claim, *disagrees* with it, or *other*. We feed BERT the claim text as sentence *A* and the evidence as sentence *B* (truncated if needed) separated by the [SEP] token. Finally, we use the representation of the [CLS] token as input to a single classification layer with three output nodes, added on top of BERT architecture to compute the probability for each class of stance.

Various Arabic BERT-based models were released recently Antoun et al. (2020); Safaya et al. (2020); Lan et al. (2020); Inoue et al. (2021); Abdul-Mageed et al. (2021); we opted to choose ARBERT Abdul-Mageed et al. (2021) as it was shown to achieve better performance on most of the stance datasets adopted in our work Alhindi et al. (2021). All models were trained with a maximum of 25 epochs where 5 was set as an early stopping threshold. We tuned our models by adopting three variant learning rates (1e-5, 2e-5, 3e-5). The sequence length and batch size were set to 512 and 16, respectively.

Preprocessing

We processed all the textual content by removing non-Arabic text, special characters, URLs, diacritics, and emojis from the tweets. For STANCEOSAURUS, we extended the tweets with their context as suggested by the authors Zheng et al. (2022) who showed that extending the tweets with parent tweet text and/or embedded article titles can improve the performance of the stance models.[7]

Loss functions

[7] We used the context extracted and shared by the authors.

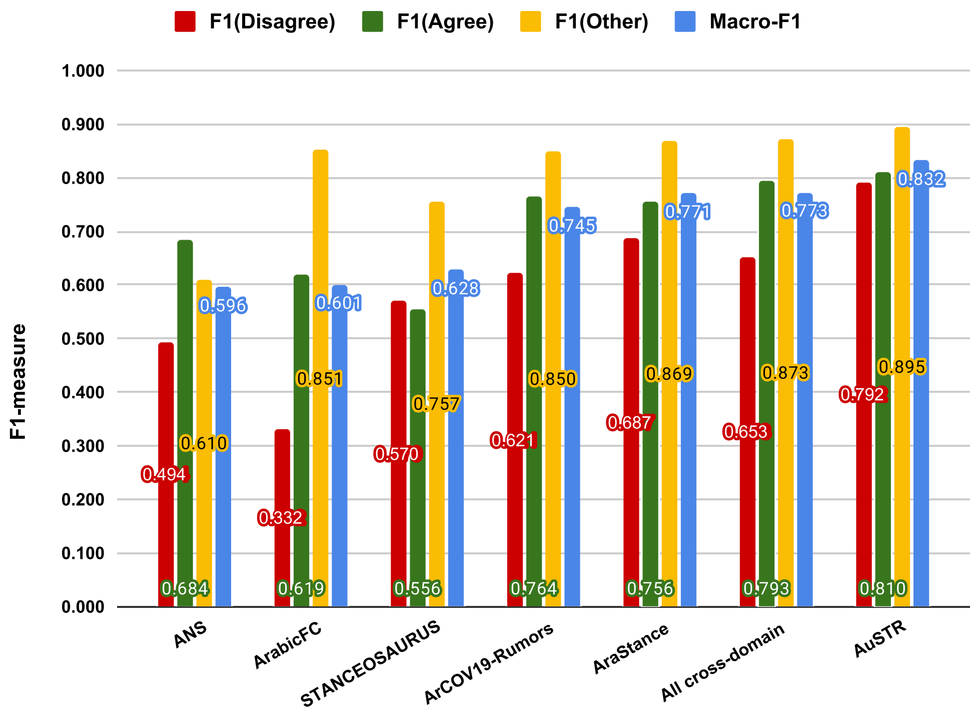


Fig. 6 The performance of models trained using *cross-domain* vs. *in-domain* datasets

We adopted the cross-entropy (*CE*) loss in all our experiments. However, due the imbalanced class distribution, we also experimented with the weighted cross-entropy (*WCE*) loss, and class-balanced focal (*CBF*) loss Cui et al. (2019) adopted by Baheti et al. (2021) and Zheng et al. (2022) to mitigate the issue for stance detection. For *CBF*, we set the hyper-parameters β and γ to 0.9999 and 1.0, respectively, as suggested by Baheti et al. (2021).

Evaluation measures

To evaluate our models, we report the average of macro-F_1 scores across the five folds of AuSTR, in addition to average per-class F_1. Macro-F_1 is recommended to evaluate stance models Hanselowski et al. (2018) due to the class-imbalance nature of stance datasets.

7 Experimental evaluation

In this section, we present and discuss the results of our experiments that address the research questions introduced in Sect. 5.

7.1 Leveraging cross-domain datasets for training (RQ1)

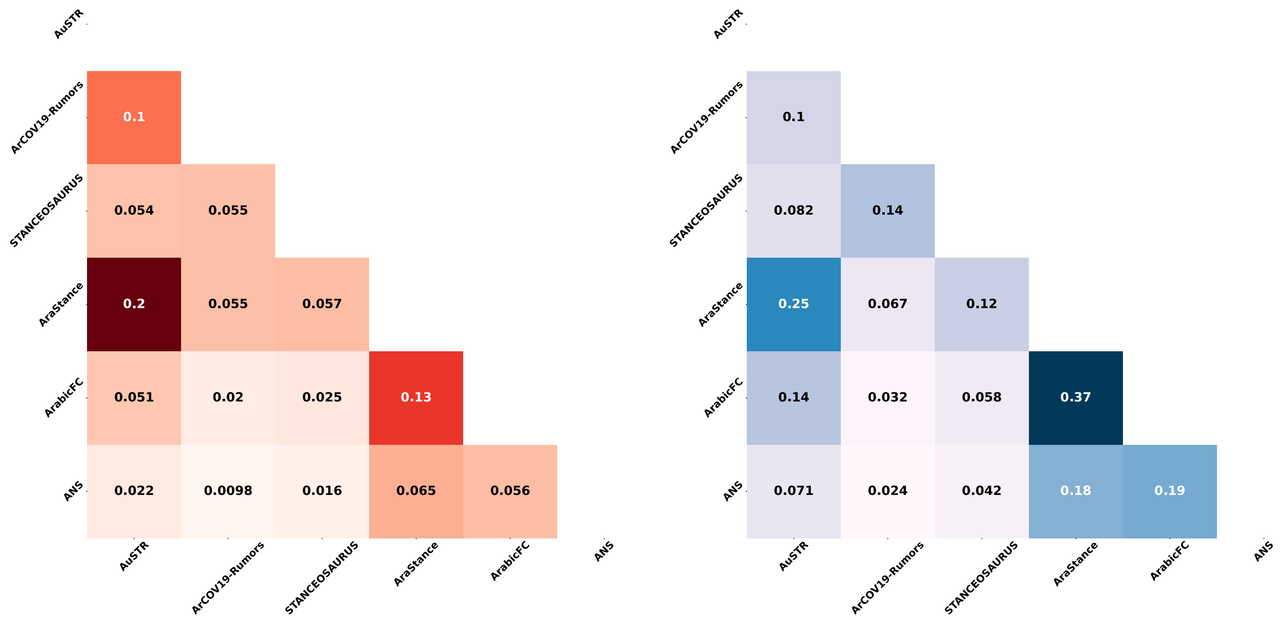
To address **RQ1**, we used the five cross-domain datasets listed earlier for training. For each of them, we train on the full cross-domain dataset, then fine-tune five stance models; each is tuned on one fold from AuSTR and tested on another fold. We report the average performance on testing on the five folds of AuSTR in Fig. 6.

The figure reveals several observations. First, the performance on the *Disagree* class is notably worse than the other two classes in four out of the five training datasets. This indicates that detecting the disagreement is generally more challenging than the agreement or irrelevance.

Second, comparing the performance across the individual cross-domain datasets, it is clear that we have two categories of performance. The first, including AraStance and ArCOV19-Rumors, is performing much better than the other one, including the remaining three datasets. Among the superior category, the model trained on AraStance exhibits the best performance.

As for the inferior category, we speculate the rationale behind their performance. We note that ArabicFC is severely imbalanced, where the *disagree* class represents only 2.86% of the data, yielding a very poor performance on that class. Moreover, it covers claims related to only one topic, which is the Syrian war, making it hard to generalize. A similar conclusion was found by previous studies that used ArabicFC Baly et al. (2018); Alhindi et al. (2021). As for ANS, evidence was manually/artificially crafted, which is not as realistic as tweets from authorities. As for STANCEOSAURUS, it covers tweets relevant to only 22 claims.

As for the superior category, we observe that AraStance and ArCOV19-Rumors achieved the highest F_1 on the *disagree* class compared to the other cross-domain datasets. ArCOV19-Rumors covers 138 COVID-19 claims in several topical categories. AraStance covers 910 claims, which are



(a)

(b)

Fig. 7 Dataset pairwise similarity using **a** debunking contexts, and **b** overall contexts

extracted from three fact-checking websites, covering multiple domains and Arab countries, similar to AuSTR, and the evidence is represented in articles written by journalists, not manually crafted. To further investigate their performance, we manually examined 20% of AraStance and ArCOV19-Rumors *disagreeing* training pairs. We found that about 68% and 59% of the examined examples of AraStance and ArCOV19-Rumors, respectively, share common debunking keywords, such as "*rumors*," "*not true*," "*denied*," and "*fake*;" similar keywords appear in some *disagreeing* tweets of AuSTR.

To further investigate the relation between the datasets and the performance of the corresponding models, we analyzed the lexical similarity between the datasets. We first constructed a 2-gram vector representation for each dataset (including AuSTR) using the preprossessed context[8](excluding the claims), then we performed a pairwise cosine similarity between the vectors to get insights about the similarity between the corresponding datasets. Figure 7a, b presents heatmaps of similarity between the debunking contexts and overall contexts of the datasets, respectively. It is clear that the performance of the cross-domain models is strongly related to the dataset similarities. In particular, AraStance has the highest similarity with AuSTR on debunking context (0.20) and overall context (0.25), respectively. That resulted in the best performing cross-domain model achieving a macro-F_1 of 0.771 and F_1 (*disagree*) of 0.687. Moreover, ArCOV19-Rumors has the second highest similarity with AuSTR on debunking context (0.10) and the second best performing cross-domain model achieving F_1(*disagree*) of 0.621. It is worth noting that although ArabicFC has the second highest similarity on the overall context, the model trained on it did not perform well especially on the *disagree* class, with F_1 of 0.332, due to the severe imbalance as mentioned earlier.

In summary, we found that AraStance is the best existing stance dataset for training a model for the task, as it covers a large number of fact-checked claims spanning multiple Arabic countries and topics compared to the other datasets. To answer **RQ1**, we conclude that some cross-domain stance datasets are somewhat useful for detecting the stance of authorities. However, motivated by the findings of Ng and Carley (2022) who highlighted the potential benefit of aggregating datasets to enhance the stance detection, we were encouraged to conduct our subsequent experiments, in which we combine all cross-domain datasets for training.

7.2 Combining cross-domain datasets for training (RQ2)

To address **RQ2**, we combined all cross-domain datasets and adopted the same setup mentioned previously, where we tune and test on AuSTR folds.

[8] For articles, we considered only the first two sentences.

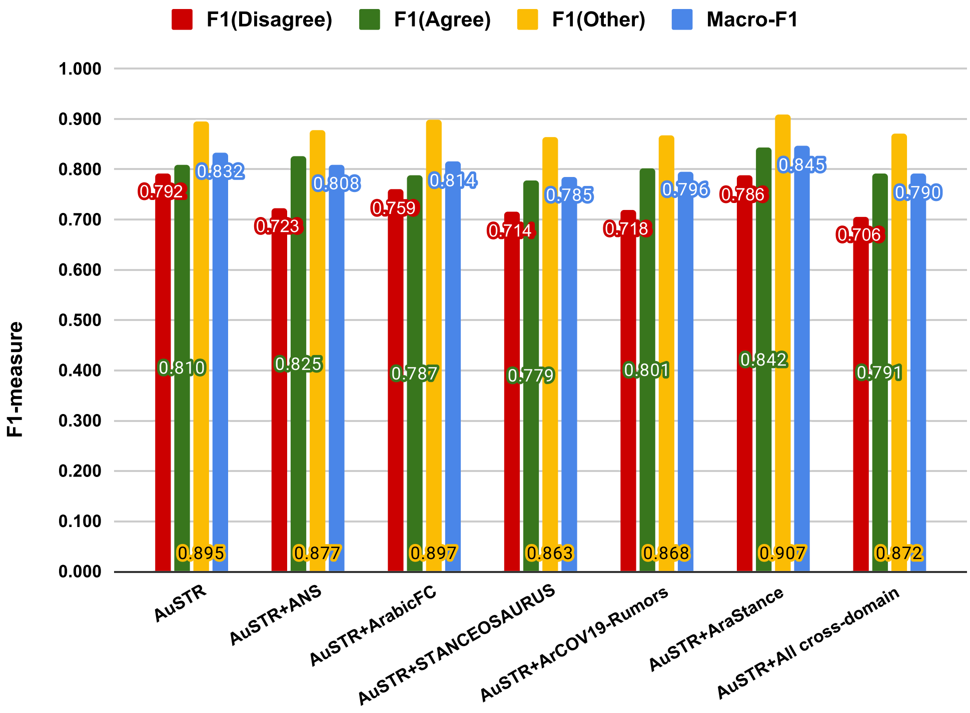


Fig. 8 Performance of models trained using *in-domain* vs. *in-domain-augmented* data

As presented in Fig. 6, we note that, overall, the combined model achieved a *very slightly* better performance in terms of macro-F_1 over the best individual model, i.e., the model trained with AraStance only. However, considering the individual classes, it exhibited the best performance for the *agree* class with a big margin compared to AraStance model, but it fell short for the *disagree* class. We speculate the reason is that some of the datasets, namely ANS and ArabicFC, achieved low performance on the *disagree* class, thus when combined with other datasets it affected negatively the overall performance on the same class.

Finally, we observe that there is a clear discrepancy in the performance across different classes; considering the combined model, F_1(*agree*) is 0.793, while F_1(*disagree*) is 0.653. Moreover, it is clear that detecting the *disagree* stance is still challenging, for which we expect to benefit from introducing our in-domain data. We believe that one of the major reasons behind such results is the imbalanced nature of the combined data, where only 14.24% are *disagree* examples vs. 27.66% *agree* examples.

To answer **RQ2**, we found that combining all cross-domain datasets can slightly improve the overall performance compared to the best performing individual model (AraStance), but could not beat it on detecting debunking tweets.

7.3 Introducing in-domain data for training (RQ3)

To address **RQ3**, we first trained a stance model with in-domain data only, i.e., AuSTR. We then trained a model with in-domain data augmented with each of the cross-domain datasets separately and also with all cross-domain datasets combined.

As expected, the model trained with AuSTR only outperforms all models trained with cross-domain datasets across all evaluation measures, as shown in Fig. 6. More specifically, it outperforms their best (i.e., the model trained with AraStance) by 15.3%, 7.1%, and 7.9% in F_1 (*disagree*), F_1(*agree*), and macro-F_1 , respectively, showing a clear need to in-domain data.

What if we augment AuSTR with the cross-domain datasets in training? Fig. 8 illustrates that effect. For every single cross-domain dataset, when augmented with AuSTR, the resulted model outperforms the model trained only on the cross-domain data by a big margin, ranging from 6.8–35.6% in macro-F_1. This re-emphasizes the effect of in-domain data. However, only the model trained on AuSTR+AraStance was able to outperform the AuSTR-only model in macro-F_1 and F_1(*agree*), but not F_1(*disagree*). It turned out that augmenting AuSTR with AraStance made the *disagree* class minority, constituting only 13.3% of the training examples compared to 24.3% of AuSTR training examples, which negatively affects the performance on that class.

Contrary to the results presented in Fig. 6, augmenting AuSTR with all cross-domain datasets achieved the lowest macro-F_1 compared to augmenting AuSTR with individual cross-domain datasets. In fact, the combined training data become clearly dominated with the cross-domain data

Table 6 Training with different loss functions. Boldfaced and underlined numbers are the best and second best, respectively, per measure

Training data	Loss function	F_1(D)	F_1(A)	F_1(O)	m-F_1
AuSTR only	*CE*	**0.792**	0.810	0.895	0.832
	WCE	0.725	0.763	0.866	0.785
	CBF	0.780	**0.844**	0.904	0.843
AuSTR+AraStance	*CE*	0.786	0.842	**0.907**	**0.845**
	WCE	0.756	0.794	0.885	0.812
	CBF	0.756	0.826	0.895	0.826

(24,313 vs. 811 examples), which leads to negligible effect of the in-domain data.

To answer **RQ3**, we conclude that in-domain data are needed for better detecting the stance of authorities. Moreover, augmenting AuSTR with AraStance improved the overall performance but at the expense of degrading the performance on detecting debunking tweets, which, we argue, is more crucial for the task.

7.4 Addressing the class-imbalance problem (RQ4)

To address **RQ4**, we selected the best two models presented in Fig. 8, namely the one trained with AuSTR only and the one trained with AuSTR augmented with AraStance. We then fine-tuned the stance models with the same previous setup but with two other loss functions, *WCE* and *CBF*, as described in Section 6.

As presented in Table 6, we observe that adopting *WCE* loss function could not improve the performance of the models compared to adopting *CE*. However, for the model trained with AuSTR, adopting *CBF* notably improved the performance over *CE* with about 4.2% on the *agree* class, which is the minority class in AuSTR data. However, it slightly degraded the performance on the *disagree* class. Overall, it improved macro-F_1 performance getting it closer to the performance of the model trained on AuSTR augmented with AraStance (0.843 vs. 0.845).

Surprisingly, that positive effect of *CBF* was not extended to the model trained on AuSTR augmented with AraStance; in fact, the performance degraded in all measures. We will leave the investigation of such result to future work.

To answer **RQ4**, we conclude that adopting *CBF* in addition to training on AuSTR solely is on bar with the model trained on both AuSTR and AraStance, nullifying the need for augmenting AuSTR with any cross-domain data for training.

8 Discussion

In this section, we discuss our evaluation results in terms of failure cases (Sect. 8.1) and limitations (Sect. 8.2).

8.1 Failure analysis

We conducted a detailed error analysis on the 113 examples (constituting 14% of the data) that failed to be predicted correctly by the model trained with AuSTR and adopting *CBF* loss. We categorize the reasons behind these errors based on a thorough examination of the failed pairs. We found that the failures can be attributed to six main reasons which we discuss below. Some failed examples are presented in Table 7.

1. **Implicit stance:** When an authority indirectly *agree* or *disagree* with the rumor. For example, P_1 is an example of a rumor about the infection of Mahmoud Al-Khatib, the director of Al-Ahly Egyptian football club, with COVID-19, and an authority tweet implicitly debunking the rumor mentioning that he is attending the training session of the team in the stadium. This failure type is the cause of 30.09% of all failures, which motivates the need to address this challenge using stance models that take this into consideration.
2. **Writing style:** Where an authority is speaking about herself, e.g., P_2. Based on our examination, 12.39% of the failures are due to this reason.
3. **Misleading debunking keywords:** When an authority is either debunking another rumor that is relevant to the topic of the target rumor, or just including some debunking keywords in his tweets even when supporting a rumor. For example, in P_3, the authority tweet mentions that the "information being posted on it today is false," although it is *agreeing* with the rumor. We found that this constitutes 10.62% of the failures.
4. **Misleading relevant keywords:** When an authority posts tweets relevant to the topic of the rumor, the model may fail to predict the stance correctly, e.g., in P_4. This constitutes 25.66% of the failed examples.
5. **Lack of context**: When an authority debunks or supports a rumor by an announcement embedded in an image or a video, e.g, in P_5. This motivates the need to consider the tweet multi-modality Jing et al. (2023); Albalawi et al. (2023) at the processing step. Moreover, some rumors may need additional context in order to be considered relevant to the authority tweet. We observed that 6.19% of the failures are of this type.
6. **Arabic MSA by authorities vs. dialects by normal users:** As opposed to English, working with Arabic language is very challenging as different dialects, i.e., informal languages, are used in different Arabic countries Abdelali et al. (2021). These dialects may have different vocabulary than the Modern Standard Arabic (MSA) which is usually used in formal communications Mubarak and Darwish (2014). Authority tweets are usually in formal language and written in MSA Arabic, while normal users may use their informal Arabic with

Table 7 Sample examples failed to be predicted correctly by our best model. Failure types are implicit stance, writing style, misleading debunking keywords, misleading relevant keywords, lack of context, and non-MSA Arabic in order

[Pair] Rumor tweet [Post date]	[Gold stance] Authority tweet [Post date]
[P_1] **@USER**: Mahmoud Al-Khatib was infected with Corona! Is the Al-Ahly administration still insisting on completing the league? Or will it change its mind after Al-Khatib was infected... [24-06-2020]	**[Disagree]** **@AlAhlyTV**: Captain Mahmoud Al-Khatib is watching our morning team's training session at the Tetch Stadium. [25-06-2020]
[P_2] **@USER**: On an official visit of 4 days. Commerce Minister Majid bin Abdullah Al-Kassabi heads a Saudi government delegation to the Kingdom of Morocco to discuss strengthening trade and investment relations. With the participation of officials from the government sector for 12 government agencies and representatives of the private sector for more than 60 Saudi companies. [03-10-2022]	**[Agree]** **@malkassabi**: Today, I had the pleasure of meeting with the Moroccan Prime Minister, Aziz Akhannouch, and we discussed strengthening our economic and commercial cooperation to meet the aspirations of the leadership of our two countries and our two brotherly peoples. [04-10-2022]
[P_3] **@USER**: Hacking the account of the Libyan Ministry of Foreign Affairs on Twitter.[22-12-2022]	**[Agree]** **@USEmbassyLibya**: The US Embassy understands that the Twitter account of the Libyan Ministry of Foreign Affairs has been hacked, and we confirm that the information being posted on it today is false. [20-12-2022]
[P_4] **@USER**: A railway network to connect the port of Sohar in the Sultanate of Oman with the city of Abu Dhabi in the UAE. [15-10-2022]	**[Other]** **@Etihad_Rail**: Etihad Rail has made significant progress in expanding the network by successfully connecting the emirates of Sharjah and Ras Al Khaimah to the main line of the UAE National Rail Network. With this achievement, the network will extend from Sharjah and Ras Al Khaimah to Al Ghuwaifat. [12-10-2022]
[P_5] **@USER**: World Cup 2022: Morocco officially protests the arbitration in the semi-finals against France. [15-12-2022]	**[Agree]** **@FRMFOFFICIEL**: Announcement from the Royal Moroccan Football Federation [Embedded image with the content of the announcement]. [15-12-2022]
[P_6] **@USER**: The first person to have monkeypox in Egypt is 39 old.. we need two nuclear bombs to close the game. [09-12-2022]	**[Agree]** **@mohpegypt**: The Ministry of Health and Population announces a positive case of monkeypox virus (Mpox) for a 39-year-old person, taking preventive measures against the infected person and his close contacts, and transferring the patient to receive treatment in one of the hospitals affiliated with the Ministry... [08-12-2022]

variant dialects, e.g, in P_6, which make detecting the stance more challenging.

We also observed other reasons, such as having multiple claims in the same tweet, which is causing the stance model to predict the authority tweet as *other*. Moreover, we noticed that some failures can be attributed to one or more of the reasons mentioned above. These challenges motivate further work on tweet preprocessing to consider embedded content within the tweets, and the need to propose stance models specific for the task.

8.2 Limitations of our study

The limitations of our work are related to both our data and the adopted stance models. We discuss these limitations below.

Data

For a portion of our data, we adopted a semi-automated approach, where we collected the *disagree* pairs starting from a collection of tweets containing debunking

keywords. Although most of the debunking tweets automatically collected where just used as pointers to collect implicit debunking tweets, some were already posted by authorities themselves and hence were considered as part of our data. This may cause some kind of bias towards these keywords. Moreover, although AuSTR with its relatively small size yielded good performance, we believe enlarging the data with more rumors covering more topics can help the models generalize better on new emerging rumors.

Stance models

In our work, we adopted a BERT-based stance model, but we did not experiment with other models, e.g., Hardalov et al. (2021) which might improve the performance we achieved. Moreover, we only experimented with ARBERT Abdul-Mageed et al. (2021) as it showed to perform well for Arabic stance detection on most of our adopted cross-domain datasets Alhindi et al. (2021); however, we did not experiment with other Arabic BERT models Abu Farha and Magdy (2021).

9 Conclusion

In this work, we introduced the task of detecting the stance of authorities towards rumors in Twitter, which can be leveraged by automated systems and fact-checkers for rumor verification. We constructed (and released) the first *Arabic* dataset, AuSTR, for that task using a language-independent approach, which we share to encourage the construction of similar datasets in other languages. Due to the relatively limited size of our dataset, we explored the adequacy of existing Arabic datasets of stance towards claims in training models for our task and the effect of augmenting our data with those datasets. Moreover, we tackled the class-imbalance issue by incorporating variant loss functions into our BERT-based stance model. Our experimental results suggest that adopting existing stance datasets is somewhat useful but clearly insufficient for detecting the stance of authorities. Moreover, when augmenting AuSTR with existing stance datasets, only the model trained with AuSTR augmented with AraStance outperformed the model trained with AuSTR solely, except on detecting the debunking tweets. However, when adopting the class-balanced focal loss instead of the cross-entropy loss, the model trained with AuSTR solely achieved comparable results to that augmented model, indicating that AuSTR solely, despite the limited size, can be sufficient for detecting the stance of authorities.

Finally, out of our extensive failure analysis, we recommend further work on tweet preprocessing to consider context expansion, and exploring other stance models that can detect the implicit stance and take the authorities writing style into consideration. Since our study focused on Arabic data, examining the task in other languages is clearly a potential path for future work.

Acknowledgements The work of Fatima Haouari was supported by GSRA grant# GSRA6-1-0611-19074 from the Qatar National Research Fund. The work of Tamer Elsayed was made possible by NPRP grant# NPRP 11 S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. Open Access funding provided by the Qatar National Library.

Author contributions FH designed the data construction approach, collected the data, developed the annotation guidelines, supervised the annotation process, designed the research questions, designed the experiments, developed and implemented the system, conducted the experiments, evaluated the performance, analyzed the results, and wrote the paper. TE supervised the data construction process, designed the research questions, designed the experiments, analyzed the results, and writing—review and editing.

Funding Open Access funding provided by the Qatar National Library.

Data availability Our data can be downloaded from our GitHub repository[1].

Declarations

Conflict of interest The authors declare they have no competing interests.

References

Abdelali A, Mubarak H, Samih Y, Hassan S, Darwish K (2021) Qadi: Arabic dialect identification in the wild. In: Proceedings of the sixth Arabic natural language processing workshop, pp 1–10

Abdul-Mageed M, Elmadany A, *et al* (2021) 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 (Vol 1: Long Papers), pp 7088–7105

Abu Farha I, Magdy W (2021) Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In: Proceedings of the sixth Arabic natural language processing workshop, pp 21–31. Association for computational linguistics, Kyiv, Ukraine (Virtual)

Albalawi RM, Jamal AT, Khadidos AO, Alhothali AM (2023) Multimodal Arabic rumors detection. IEEEIEEEIEEE Access 11:9716–9730

Alhindi T, Alabdulkarim A, Alshehri A, Abdul-Mageed M, Nakov P (2021) AraStance: A Multi-Country and Multi-Domain Dataset of Arabic Stance Detection for Fact Checking. NLP4IF 2021, 57

Ali ZS, Mansour W, Elsayed T, Al-Ali A (2021) AraFacts: the first large arabic dataset of naturally occurring claims. In: Proceedings of the sixth Arabic natural language processing workshop, pp 231–236

Alqurashi S, Hamoui B, Alashaikh A, Alhindi A, Alanazi E (2021) Eating garlic prevents COVID-19 infection: detecting misinformation on the arabic content of twitter. arXiv preprint arXiv:2101.05626

Alshaabi T, Dewhurst DR, Minot JR, Arnold MV, Adams JL, Danforth CM, Dodds PS (2020) The growing echo chamber of social media: measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009-2020. CoRR **abs/2003.03667**

Althabiti S, Alsalka MA, Atwell E (2022) Detecting Arabic fake news on social media using sarcasm and hate speech in comments. Int J Islam Appl Comput Sci Technol 10(4):28–36

Alturayeif NS, Luqman HA, Ahmed MAK (2022) Mawqif: a multi-label Arabic dataset for target-specific stance detection. In: Proceedings of the the seventh Arabic natural language processing Workshop (WANLP), pp 174–184. Association for computational linguistics, Abu Dhabi, United Arab Emirates (Hybrid)

Al-Yahya M, Al-Khalifa H, Al-Baity H, AlSaeed D, Essam A (2021) Arabic fake news detection: comparative study of neural networks and transformer-based approaches. Complexity 2021:1–10

Antoun W, Baly F, Hajj H (2020) AraBERT: transformer-based model for arabic language understanding. In: LREC 2020 workshop language resources and evaluation conference 11–16 May 2020, p 9

Baheti A, Sap M, Ritter A, Riedl M (2021) Just say no: analyzing the stance of neural dialogue generation in offensive contexts. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 4846–4862. Association for computational linguistics, Online and Punta Cana, Dominican Republic

Bai N, Meng F, Rui X, Wang Z (2022) A multi-task attention tree neural net for stance classification and rumor veracity detection. Appl Intell 53(9):10715–10725

Bai N, Meng F, Rui X, Wang Z (2022) Rumor detection based on a source-replies conversation tree convolutional neural net. Computing 104(5):1155–1171

Baly R, Mohtarami M, Glass J, Màrquez L, Moschitti A, Nakov P (2018) 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, Vol 2 (Short Papers), pp 21–27. Association for computational linguistics, New Orleans, Louisiana

Barrón-Cedeño A, Alam F, Caselli T, Da San Martino G, Elsayed T, Galassi A, Haouari F, Ruggeri F, Struß JM, Nandi RN, *et al* (2023) The CLEF-2023 checkthat! lab: checkworthiness, subjectivity, political bias, factuality, and authority. In: Advances in information retrieval: 45th European conference on information retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, pp 506–517

Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, Huang J (2020) Rumor detection on social media with bi-directional graph convolutional networks. In: Proceedings of the AAAI conference on artificial intelligence, pp 549–556

Chen L, Wei Z, Li J, Zhou B, Zhang Q, Huang XJ (2020) Modeling evolution of message interaction for rumor resolution. In: Proceedings of the 28th international conference on computational linguistics, pp 6377–6387

Choi J, Ko T, Choi Y, Byun H, Kim Ck (2021) Dynamic graph convolutional networks with attention mechanism for rumor detection on social media. Plos one 16(8):0256039

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Cui Y, Jia M, Lin TY, Song Y, Belongie S (2019) Class-balanced loss based on effective number of samples. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9260–9269. IEEE

Derczynski L, Bontcheva K, Liakata M, Procter R, Wong Sak Hoi G, Zubiaga A (2017) SemEval-2017 task 8: rumourEval: determining rumour veracity and support for rumours. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp 69–76. Association for computational linguistics, Vancouver, Canada

Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Dougrez-Lewis J, Kochkina E, Arana-Catania M, Liakata M, He Y (2022) PHEMEPlus: enriching social media rumour verification with external evidence. In: Proceedings of the fifth fact extraction and verification workshop (FEVER), pp 49–58

Elhadad MK, Li KF, Gebali F (2020) 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

Ferreira W, Vlachos A (2016) Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1163–1168. Association for computational linguistics, San Diego, California

Gorrell G, Kochkina E, Liakata M, Aker A, Zubiaga A, Bontcheva K, Derczynski L (2019) SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In: Proceedings of the 13th international workshop on semantic evaluation, pp 845–854. Association for computational linguistics, Minneapolis, Minnesota, USA

Guderlei M, Aßenmacher M (2020) Evaluating unsupervised representation learning for detecting stances of fake news. In: Proceedings of the 28th international conference on computational linguistics, pp 6339–6349. International committee on computational linguistics, Barcelona, Spain (Online)

Hanselowski A, PVS A, Schiller B, Caspelherr F, Chaudhuri D, Meyer CM, Gurevych I (2018) A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th international conference on computational linguistics, pp. 1859–1874. Association for computational linguistics, Santa Fe, New Mexico, USA

Hanselowski A, Avinesh P, Schiller B, Caspelherr F, Chaudhuri D, Meyer CM, Gurevych I (2018) A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th international conference on computational linguistics, pp 1859–1874

Haouari F, Hasanain M, Suwaileh R, Elsayed T (2021) ArCOV19-rumors: Arabic COVID-19 twitter dataset for misinformation detection. In: Proceedings of the sixth Arabic natural language processing workshop, pp 72–81

Haouari F, Elsayed T (2023) Detecting stance of authorities towards rumors in Arabic tweets: a preliminary study. In: Advances in information retrieval, pp 430–438. Springer, Cham

Haouari F, Elsayed T, Mansour W (2023) Who can verify this? Finding authorities for rumor verification in twitter. Inf Process Manag 60(4):103366

Haouari F, Sheikh Ali Z, Elsayed T (2023) Overview of the CLEF-2023 checkthat! lab task 5 on authority finding in twitter. In: Working notes of CLEF 2023–conference and labs of the evaluation forum. CLEF ’2023, Thessaloniki, Greece

Hardalov M, Arora A, Nakov P, Augenstein I (2021) Cross-domain label-adaptive stance detection. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 9011–9028

Hardalov M, Arora A, Nakov P, Augenstein I (2022) A survey on stance detection for mis-and disinformation identification. In: Findings of the association for computational linguistics: NAACL 2022, pp 1259–1277

Hasanain M, Haouari F, Suwaileh R, Ali ZS, Hamdan B, Elsayed T, Barrón-Cedeno A, Da San Martino G, Nakov P (2020)

Overview of checkthat! 2020 Arabic: automatic identification and verification of claims in social media. In: CLEF

Hu X, Guo Z, Chen J, Wen L, Yu PS (2023) Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, pp 2901–2912

Inoue G, Alhafni B, Baimukan N, Bouamor H, Habash N (2021) 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

Jing J, Wu H, Sun J, Fang X, Zhang H (2023) Multimodal fake news detection via progressive fusion networks. Inf Process Manag 60(1):103120

Khouja J (2020) 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

Kumar S, Carley K (2019) 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

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Lan W, Chen Y, Xu W, Ritter A (2020) 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

Li Y, Scarton C (2020) Revisiting rumour stance classification: dealing with imbalanced data. In: Proceedings of the 3rd international workshop on rumours and deception in social media (RDSM), pp 38–44. Association for computational linguistics, Barcelona, Spain (Online)

Liu Y, Wu YFB (2018) 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

Liu Z, Xiong C, Sun M, Liu Z (2020) Fine-grained fact verification with kernel graph attention network. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 7342–7351. Association for computational linguistics, Online

Ma J, Gao W (2020) Debunking rumors on twitter with tree transformer. In: Proceedings of the 28th international conference on computational linguistics, pp 5455–5466. International committee on computational linguistics, Barcelona, Spain (Online)

Mahlous AR, Al-Laith A (2021) Fake news detection in Arabic tweets during the COVID-19 pandemic. Int J Adv Comput Sci Appl 12(6):778–788

Ma J, Gao W, Wong KF (2018) Rumor detection on twitter with tree-structured recursive neural networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Vol 1: Long Papers), pp 1980–1989

Mubarak H, Darwish K (2014) Using twitter to collect a multi-dialectal corpus of Arabic. In: Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP), pp 1–7

Mu Y, Niu P, Aletras N (2022) Identifying and characterizing active citizens who refute misinformation in social media. In: 14th ACM web science conference 2022, pp 401–410

Ng LHX, Carley KM (2022) Is my stance the same as your stance? A cross validation study of stance detection datasets. Inf Process Manag 59(6):103070

Nyhan B, Reifler J (2015) Displacing misinformation about events: an experimental test of causal corrections. J Exp Polit Sci 2(1):81–93

Pomerleau D, Rao D (2017) Fake news challenge stage 1 (fnc-i): stance detection http://www.fakenewschallenge.org/#fnc1-scoring

Roy S, Bhanu M, Saxena S, Dandapat S, Chandra J (2022) gDART: improving rumor verification in social media with discrete attention representations. Inf Process Manag 59(3):102927

Safaya A, Abdullatif M, Yuret D (2020) 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)

Sawan A, Thaher T, Abu-el-rub N (2021) 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

Shaar S, Hasanain M, Hamdan B, Ali ZS, Haouari F, Nikolov A, Kutlu M, Kartal YS, Alam F, Da San Martino G, et al (2021) Overview of the CLEF-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates

Sheikh Ali Z, Mansour W, Haouari F, Hasanain M, Elsayed T, Al-Ali A (2023) Tahaqqaq: a real-time system for assisting twitter users in arabic claim verification. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval

Si J, Zhou D, Li T, Shi X, He Y (2021) Topic-aware evidence reasoning and stance-aware aggregation for fact verification. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Vol 1: Long Papers), pp 1612–1622. Association for computational linguistics, Online

Slovikovskaya V, Attardi G (2020) Transfer learning from transformers to fake news challenge stance detection (FNC-1) task. In: Proceedings of the twelfth language resources and evaluation conference, pp 1211–1218. European language resources association, Marseille, France

Song C, Shu K, Wu B (2021) Temporally evolving graph neural network for fake news detection. Inf Process Manag 58(6):102712

Thorne J, Vlachos A, Christodoulopoulos C, Mittal A (2018) FEVER: a large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, Vol 1 (Long Papers), pp 809–819. Association for computational linguistics, New Orleans, Louisiana

Vo N, Lee K (2018) The rise of guardians: fact-checking url recommendation to combat fake news. The 41st international ACM SIGIR conference on research & development in information retrieval. SIGIR ’18. Association for computing machinery, New York, NY, USA, pp 275–284

Vo N, Lee K (2019) Learning from fact-checkers: analysis and generation of fact-checking language. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 335–344

Vo N, Lee K (2020) Where are the facts? Searching for fact-checked information to alleviate the spread of fake news. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 7717–7731

Vo N, Lee K (2020) Standing on the shoulders of guardians: novel methodologies to combat fake news. In: Disinformation, misinformation, and fake news in social media: emerging research challenges and opportunities, pp 183–210

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151

Wu L, Rao Y, Jin H, Nazir A, Sun L (2019) 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

You D, Vo N, Lee K, LIU Q (2019) Attributed multi-relational attention network for fact-checking URL recommendation. In: Proceedings of the 28th ACM international conference on information and knowledge management. CIKM ’19, pp 1471–1480. Association for computing machinery, New York, NY, USA

Yu J, Jiang J, Khoo LMS, Chieu HL, Xia R (2020) 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

Zheng J, Baheti A, Naous T, Xu W, Ritter A (2022) Stanceosaurus: classifying stance towards multicultural misinformation. In: Proceedings of the 2022 conference on empirical methods in natural language processing, pp 2132–2151. Association for computational linguistics, Abu Dhabi, United Arab Emirates

Zhou J, Han X, Yang C, Liu Z, Wang L, Li C, Sun M (2019) GEAR: graph-based evidence aggregating and reasoning for fact verification. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 892–901. Association for computational linguistics, Florence, Italy

Zubiaga A, Liakata M, Procter R, Hoi Wong Sak G, Tolmie P (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one 11(3):0150989

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.