AuRED: Enabling Arabic Rumor Verification using Evidence from Authorities over Twitter

Fatima Haouari
Qatar University, Qatar
200159617@qu.edu.qa

Tamer Elsayed
Qatar University, Qatar
telsayed@qu.edu.qa

Reem Suwaileh
HBKU, Qatar
rsuwaileh@hbku.edu.qa

Abstract

Diverging from the trend of the previous rumor verification studies, we introduce the new task of rumor verification using *evidence* that are exclusively captured *from authorities*, i.e., entities holding the right and knowledge to verify corresponding information. To enable research on this task for *Arabic low-resourced language*, we construct and release the *first* Authority-Rumor-Evidence Dataset (AuRED). The dataset comprises 160 rumors expressed in tweets and 692 Twitter timelines of authorities containing about 34k tweets. Additionally, we explore how existing evidence retrieval and claim verification models for fact-checking perform on our task under both the *cross-lingual zero-shot* and *in-domain fine-tuning* setups. Our experiments show that although evidence retrieval models perform relatively well on the task establishing strong baselines, there is still a big room for improvement. However, existing claim verification models perform poorly on the task no matter how good the retrieval performance is. The results also show that stance detection can be useful for evidence retrieval. Moreover, existing fact-checking datasets showed a potential in transfer learning to our task, however, further investigation using different datasets and setups is required.

1 Introduction

The spread of rumors and fake news on social media causes anxiety and panic in communities, forming persistent challenges for platforms, policymakers, and researchers. To address this, several rumor verification studies on social media incorporate the propagation networks as a key source of evidence. They either utilize the stance of replies (Kumar and Carley, 2019; Yu et al., 2020; Bai et al., 2023), the structure of the replies (Ma et al., 2018; Bian et al., 2020; Song et al., 2021), or the users' metadata (Liu and Wu, 2018). On the other hand, evidence are extracted from the Web to augment signals from the propagation networks (Dougrez-Lewis et al., 2022; Hu et al., 2023). However, studies on *Arabic* rumor verification incorporating evidence are scarce. Haouari et al. (2021) and Althabiti et al. (2022) exploited the tweet replies, while Albalawi et al. (2023) leveraged the images and videos embedded in the rumor tweet.

Authorities (i.e., entities having the real knowledge or power to verify or deny a specific rumor (Haouari et al., 2023; Haouari and Elsayed, 2023)) can also be a valuable source of evidence that augments other sources for verifying rumors, either by automated verification systems or more specifically by human fact-checkers. Detecting the stance of authorities towards rumors in Twitter[1] was indeed introduced recently as a potential signal for better rumor verification (Haouari and Elsayed, 2023, 2024). However, to the best of our knowledge, no study to date has explored the incorporation of *evidence tweets* retrieved from the timelines of authorities for rumor verification over social media in general and for *Arabic* rumor verification in particular. Additionally, there is no available dataset for that task to support such research.

To bridge this gap, in this paper, we introduce the problem of *rumor verification using evidence from authorities over Twitter* and a *dataset* that enables research tackling that problem. The problem is defined as follows: given a rumor expressed in a tweet and a set of Twitter accounts of authorities for that rumor, the system should retrieve evidence tweets posted by any of those authorities. Based on the retrieved evidence, the system should determine if the rumor is supported, refuted, or unverifiable. Figure 1 illustrates the setup of the problem.

To facilitate the research on this task, we introduce **AuRED**, the first **Au**thority-**R**umor-**E**vidence **D**ataset. AuRED covers 160 *Arabic* rumors an-

[1]"Twitter" is the former name of "X," however we will use "Twitter" for clarity.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 27–41
August 16, 2024 ©2024 Association for Computational Linguistics

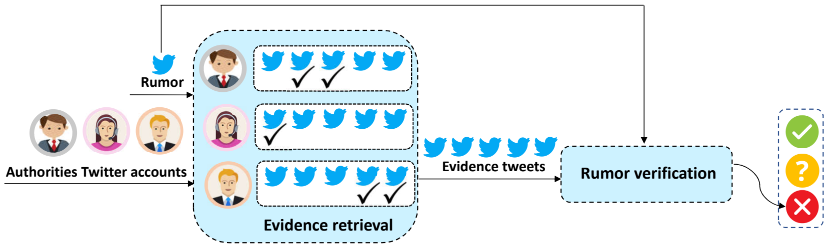

Figure 1: Rumor verification using evidence from authorities over Twitter pipeline.

notated with tweet-level evidence from their corresponding 692 authority timelines, comprising about 34k annotated tweets in total. The dataset was constructed by annotating a set of rumors, selected from two existing datasets (Haouari et al., 2023; Haouari and Elsayed, 2024), following two main steps (1) finding authorities that can help verify the rumors (Haouari et al., 2023), and (2) collecting the timelines of those authorities, and annotating those timelines to find evidence tweets.

Our contribution in this work is five-fold:

- We propose the new task of *rumor verification using evidence from authorities over Twitter*.
- We introduce AuRED,[2] the first *Arabic* public dataset for the task.
- We present benchmarking results on AuRED, and release our source code for reproducibility and facilitating research on the task.[3]
- We explore how existing evidence retrieval and claim verification models, that are originally proposed for fact-checking, perform on our task under both the *cross-lingual zero-shot* and *in-domain fine-tuning* setups.
- We investigate the usefulness of detecting stance of authorities toward rumors for the evidence retrieval subtask.

The remainder of the paper is organized as follows. We review the literature in Section 2 and formally define our task in Section 3. In Section 4, we discuss our dataset construction approach. Our experimental design and setup are presented in Sections 5 and 6, respectively. We analyze our experimental results and answer our research questions in Section 7. We conclude and suggest some future directions in Section 8. Finally, we discuss the limitations of our work in Section 9.

[2]https://github.com/Fatima-Haouari/AuRED

[3]Our released resources are presented in Appendix B.4.

2 Related Work

In this section, we review previous studies on rumor verification in social media and fact-checking.

Rumor Verification in Social media. There exists a considerable body of literature on rumor verification in social media (Ma et al., 2018; Kumar and Carley, 2019; Yu et al., 2020; Choi et al., 2021; Bai et al., 2022a). The majority of prior research has leveraged the propagation networks such as the structure of replies (Ma et al., 2018; Bian et al., 2020; Haouari et al., 2021; Bai et al., 2022b), stance of replies (Zubiaga et al., 2016; Derczynski et al., 2017), or retweeters metadata (Liu and Wu, 2018). In addition to the propagation networks, incorporating evidence from the Web was proposed by Dougrez-Lewis et al. (2022) and Hu et al. (2023). Moreover, Haouari and Elsayed (2023, 2024) proposed recently leveraging the stance of authority tweets towards rumors.

Although there are many studies, the research in *Arabic* Rumor verification remains limited. Previous studies have almost exclusively utilized the tweet textual content for verification (Elhadad et al., 2021; Mahlous and Al-Laith, 2021; Al-Yahya et al., 2021; Alqurashi et al., 2021; Sawan et al., 2021). Recently, Haouari et al. (2021) leveraged the replies structure, Althabiti et al. (2022) incorporated the detected sarcasm and hate speech in the replies, while Albalawi et al. (2023) exploited the images and videos embedded in the rumor tweet. Differently, in our work we propose using the evidence tweets retrieved from the authority timelines.

Most of the existing datasets (refer to Appendix A) focus on using the propagation networks as evidence, while the majority of *Arabic* Rumor verification datasets do not incorporate any external evidence. Compared to existing datasets, AuRED incorporates evidence from authority timelines.

Fact-Checking. Claim verification using evidence from Wikipedia was introduced as part of

FEVER shared task by Thorne et al. (2018a). The task is a pipeline of three subtasks namely documents retrieval, evidence selection, and claim verification. A plethora of studies addressed the task contributing either to the evidence retrieval or claim verification or both. For evidence retrieval, existing studies either adopted neural ranking models (Hanselowski et al., 2018; Zhou et al., 2019; Nie et al., 2019a,b) or pre-trained models (Liu et al., 2020; Jiang et al., 2021; DeHaven and Scott, 2023). Recent studies, exploited pre-trained models but with variant loss functions and some additional enhancements. Some addressed the task as a binary classification task (Zhong et al., 2020; Si et al., 2021; Jiang et al., 2021; DeHaven and Scott, 2023), some proposed a pairwise ranking model (Soleimani et al., 2020; Liu et al., 2020), while others explored distance-based loss functions (Bekoulis et al., 2021). For the claim verification task, most of the studies formulated it as a graph-based reasoning task (Zhou et al., 2019; Liu et al., 2020; Zhong et al., 2020; Park et al., 2022). Others, proposed incorporating the topic and implicit stance of evidence using the capsule network (Si et al., 2021; Ma et al., 2022), or multi-level attention (Kruengkrai et al., 2021).

In our work, we consider evidence retrieval from authority timelines and rumor verification using evidence from authorities similar to the evidence selection and the claim verification for fact-checking tasks respectively. Moreover, we investigate the knowledge transfer ability of existing fact-checking datasets to our task. One of such datasets is FEVER (Thorne et al., 2018a), an English fact checking dataset containing 185,445 claims, and their relevant evidence sentences from Wikipedia.

3 Task definition

We propose the task of *Rumor Verification using Evidence from Authorities* with two subtasks:

- **Evidence Retrieval**: Given a rumor expressed in a tweet and a set of authorities for that rumor, the system should retrieve *evidence tweets* posted by any of those authorities. An evidence tweet is a tweet that can be further used to detect the veracity of the rumor. The set of authorities has one or more authority Twitter accounts, represented by a list of tweets from their timelines that are posted during the period surrounding the rumor.

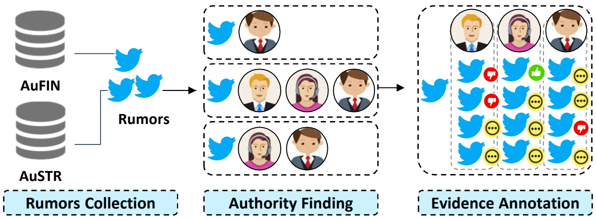

Figure 2: AuRED construction process.

- **Rumor Verification**: Based solely on the evidence tweets retrieved by the above subtask, determine if the rumor is *supported* (true), *refuted* (false), or *unverifiable* (in case not enough evidence to verify it exists).

4 AuRED Dataset

To expedite the development of automatic verification systems and to evaluate proposed models for our task, we introduce the first **Au**thority-**R**umor-**E**vidence **D**ataset (AuRED). We target Arabic as it is one of the most used languages in Twitter (Alshaabi et al., 2021), yet under-explored for rumor verification. As presented in Figure 2, the dataset was constructed by annotating a set of rumors, selected from two existing datasets (Section 4.1) following two main steps (1) finding authorities that can help verify the rumors (Section 4.2), and (2) collecting and annotating the timelines of those authorities to find evidence tweets (Section 4.3).

4.1 Rumors Collection

Due to time and budget constraints, we randomly selected 160 rumors from AuFIN (Haouari et al., 2023) and AuSTR (Haouari and Elsayed, 2024) datasets. AuFIN is an Arabic test collection for authority finding in Twitter, where each rumor is associated with its relevant authorities. AuSTR is an Arabic dataset for detecting the stance of authorities towards rumors. Given that all AuFIN rumors were collected originally from a fact-checking Website, it lacks true (i.e., confirmed) rumors as fact-checkers focus mainly on verifying false (i.e., denied) rumors, we had to get all of our 30 true rumors from AuSTR dataset. Moreover, we selected 31 false rumors from AuSTR, as each has already at least one authority tweet refuting it. In total, 99 (61.9%) of our rumors are from AuFIN while 61 (38.1%) are from AuSTR.

4.2 Authority Finding

Finding authorities in Twitter for a specific rumor was proposed recently by Haouari et al. (2023). They define an authority for a specific rumor as *an entity having the real knowledge or power to verify or deny that rumor.* For example, if the rumor is about a health issue in Iraq, then the health minister, ministry, or other leaders in health organizations in Iraq are potential authorities.

AuFIN rumors are already associated with their relevant authorities, however AuSTR rumors are only associated with an authority tweet either supporting, refuting or irrelevant to the rumor. Therefore, for AuSTR rumors, in addition to considering the authority of the associated authority tweet, we collected more authorities for each rumor following the same approach proposed by Haouari et al. (2023). Two annotators, a PhD holder and a PhD candidate, performed the task independently, then met to discuss their annotations. Only potential authorities that both annotators agreed upon during their meeting were kept in AuRED.

4.3 Evidence Annotation

In the context of this work, we consider the rumor tweet as a pointer to the period of the rumor propagation, assuming that the rumor is circulating for a few days before and/or after the time at which the tweet containing it is posted. Therefore, for evidence annotation, we limit the authority timelines to the tweets within 3 days before and after the posting time of the rumor tweet. The timelines were collected using the Academic Twitter search API which facilitates collecting user historical timelines.[4] We carried out two stages for evidence extraction:

(a) Annotation: Following our annotation guidelines, one annotator labeled *all* tweets in *all* authority timelines as *supporting*, *refuting*, or *carrying not enough info* towards the corresponding rumor tweet (constituting AuRED core dataset). To measure the quality of our data, and to have a double-annotated sample, a second annotator then labeled solely *one* authority timeline per rumor (constituting AuRED* subset). To ensure the inter-annotator consistency, we asked the annotators to ask themselves this general question: *If I was given the authority tweet, do I have a strong evidence to decide if the rumor is true (supported), false (refuted), or unverifiable (Not enough information to verify it).* At the end of this stage, we measured the data quality of AuRED* using Cohen's Kappa for inter-annotator agreement (Cohen, 1960) as 0.67, which indicates "substantial" agreement (Landis and Koch, 1977). It is worth noting, that any disagreement between the annotators was then resolved in the next stage.

(b) Resolving Disagreements: As a final step, both annotators met to discuss and resolve any disagreements in AuRED*, and hence decide the final labels. The statistics about our AuRED and AuRED* are presented in Table 1. We present examples from AuRED, our annotation challenges, and some data analysis in Appendix B.

Table 1: AuRED Statistics.

Rumors		
SUPPORTS	30	(18.75%)
REFUTES	64	(40%)
NOT ENOUGH INFO	66	(41.25%)
AuRED Authority tweets		
Authorities	692	
Average per rumor	4.33	
Authority tweets	33,705	
Median per rumor	129	
SUPPORTS	118	
REFUTES	306	
NOT ENOUGH INFO	33,281	
Videos	4,998	
Images	17,817	
AuRED* Authority tweets		
Authorities	160	
Average per rumor	1	
Authority tweets	9,755	
Median per rumor	23	
SUPPORTS	75	
REFUTES	213	
NOT ENOUGH INFO	9,467	

5 Experimental Design

Our task is closely related to the general task of *fact-checking* (Thorne et al., 2018b). In fact, it can be viewed as a special case of the fact-checking task, where evidence for verification is exclusively retrieved from *authorities* rather than from any other source, e.g., Web pages, or posts from layman users or propagation networks on social media. With a large body of existing research on the fact-checking task (Nakov et al., 2021), it is intriguing to investigate how existing evidence retrieval and claim verification models, originally designed for the general fact-checking task, perform on our specific task. Moreover, with the availability of datasets for the general task in other languages

[4] https://developer.x.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

(e.g., FEVER (Thorne et al., 2018a)), it is then intuitive to explore the potential of cross-lingual transfer learning. Accordingly, we address the following research questions:

- **RQ1**: How effective are the existing models for our task under the *cross-lingual zero-shot* setup?
- **RQ2**: How do existing models perform on our task if they are directly *fine-tuned* with AuRED?

It is worth noting that for each of the two research questions, we evaluate the performance of the models on AuRED for the two sub-tasks. Accordingly, to address both questions, we design our experiments as follows:

- **Cross-lingual Zero-shot Setup**: We study the performance of existing models on AuRED when they are fine-tuned only on English data for evidence retrieval and rumor verification, without being fine-tuned on AuRED.
- **In-domain Fine-tuning Setup**: We study the performance of existing models on AuRED when they are directly fine-tuned on AuRED.

6 Experimental Setup

In this section, we present our detailed experimental setup. We discuss our adopted evidence retrieval and rumor verification models in Sections 6.1 and 6.2, respectively. We also discuss how we evaluate those models in Section 6.3.

6.1 Evidence Retrieval Models

In addition to evaluating strong sparse and dense retrieval approaches, we selected two SOTA models (KGAT and MLA) for evidence retrieval which exhibited the best performance on FEVER test set (Park et al., 2022).[5] Moreover, we explore a model with a distance-based loss function. Finally, we adopted a stance-based approach for evidence retrieval. It is worth noting that although 49.05% of AuRED evidence tweets are multimodal, all the models we adopted in this work considers only the textual content of the tweets. In this section, we present the models and their implementation details.

1. **BM25**: One of the most successful lexical retrieval models (Jones et al., 2000). Using Pyserini (Lin et al., 2021), we constructed an index for all tweets from all authorities for a each rumor. We then retrieved, for each rumor, the top relevant authority tweets from the corresponding index.

2. **mContriever** (Izacard et al., 2021): A multilingual dense retrieval model that achieves good retrieval performance on Arabic data when further fine-tuned using MS MARCO dataset. For each rumor, we retrieved tweets that are the closest in the Contriever's embedding space using cosine similarity.

3. **KGAT** (Liu et al., 2020): A widely adopted retrieval model in fact-checking studies (Zhao et al., 2019; Park et al., 2022; Ma et al., 2022; Chen et al., 2022). It is a pairwise BERT-based model where the margin ranking loss is adopted to maximize the distance between the positive and the negative claim-sentence pairs. As suggested by the authors, the model during training was fine-tuned to maximize the distance between each positive and negative rumor-tweet-authority-tweet pairs for all authority tweets for a specific rumor. At inference, the score predicted for each rumor-tweet-authority-tweet pair is used to retrieve the top evidence tweets. We adopted the authors' implementation.[6]

4. **MLA** (Kruengkrai et al., 2021): A pointwise BERT-based binary classifier to detect evidence vs. non-evidence. The cross entropy loss was adopted. For negative examples, the authors proposed sampling M non-evidence sentences from the labelled documents and M from retrieved potentially-relevant documents, where M is twice the number of evidences. In our work, we only have the labelled documents (timelines), so we considered the number of non-evidence tweets to be 4 times the number of evidence tweets for each rumor.[7] At training and inference, rumor-tweet-authority-tweet pair are fed to a BERT-based model separated by a [SEP] token. The authors' code was adopted for our experiments.[8]

5. **TML** (Bekoulis et al., 2021): We investigate the performance when adopting the triplet

[5]The model proposed recently by DeHaven and Scott (2023) is SOTA but they adopt re-retrieval using hyperlinks in retrieved sentences to beat KGAT. Re-retrieval is not applicable in our work

[6]https://github.com/thunlp/KernelGAT

[7]Based on our preliminary experiments we found that 4 is the best considering 2, 4, 6, and 8 when fine-tuning

[8]https://github.com/nii-yamagishilab/mla

margin loss (TML), compared to the pointwise (MLA) and the pairwise (KGAT) models. This loss minimizes the pairwise distance between the rumor and the evidence, and maximizes the distance between the rumor and non-evidence. As suggested by the authors, the evidence and the non-evidence tweets are prepended with the rumor and a [SEP] token. During inference, the pairwise distance is computed between each rumor and its corresponding authority tweets (prepended by rumor [SEP]) to select the top with the lowest distance. We adopted the authors' code.[9]

6. **STAuRED**: Motivated by the task of detecting the stance of authorities (Haouari and Elsayed, 2023, 2024) as a source of evidence, we fine-tuned BERT-based stance detection model using AuRED to classify whether an authority tweet SUPPORTS, REFUTES, or NOT ENOUGH INFO. We feed BERT the rumor tweet as sentence A and the authority tweet as sentence B 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 stance class. For retrieving the top evidence tweets, we considered the sum of the softmax scores of both SUPPORTS and REFUTES labels as a reranking score.

Implementation details: For evaluation, we adopted a cross validation setup where we split our AuRED dataset into 5 folds, each containing 32 rumors ensuring balance across rumors labels. We fine-tuned the models using 3 folds and we selected the best model based on Mean Average Precision (MAP) on the dev set for each fold. We fine-tuned using 4 different learning rates [2e-5, 3e-5, 4e-5, 5e-5]. We trained all the models for 5 epochs using a batch of size 8. As our dataset contains tweets only, we adopted MARBERTv2 (Abdul-Mageed et al., 2021),[10] an Arabic BERT model pre-trained using 1 billion Arabic tweets. For the cross-lingual evidence retrieval setup, we adopted the original setup suggested by the authors, i.e., fine tuning the models with English FEVER (Thorne et al., 2018a), but we replaced the English BERT with multilingual BERT (mBERT) (Devlin et al., 2019).[11] We retrieved the top 5 evidence tweets for each rumor.

6.2 Rumor Verification Models

To have a full pipeline for both evidence retrieval and rumor verification, in our experiments we adopted both MLA and KGAT where models for both subtasks were proposed by the authors:[12]

1. **MLA** (Kruengkrai et al., 2021): It adopts multi-task learning considering the verification as the main task, and evidence retrieval as an auxiliary task where it incorporates the evidence retrieval scores through joint training. The model applies token-level attention over a claim-evidence pair, token and sentence-level self-attentions for evidence sentences. Finally, it combines all hidden states with the evidence retrieval scores at the final attention layer.

2. **KGAT** (Liu et al., 2020): A Kernel Graph Attention Network that utilizes the retrieved evidence to construct a fully connected graph and perform reasoning to verify the claims. Each node in the graph is represented using the [CLS] token of a pre-trained BERT, by feeding it a concatenation of the claim and the evidence separated by a [SEP] token.

Implementation details: During training, Both MLA and KGAT prepend the gold evidence (decided by the annotators) to the retrieved evidence, and take as input both the rumor and 5 evidence tweets. At inference time, only the retrieved evidence is considered to verify the rumors. We adopted the same cross validation setup adopted for evidence retrieval, but we fine-tuned the models based on the best Macro-F1 on the dev set.

6.3 Evaluation Scenarios and Measures

6.3.1 Evidence Retrieval

To evaluate the performance of the evidence retrieval models, we considered two sets of measures based on two scenarios as presented below:

- The **User Scenario** is the case where a human, mostly a fact-checker, is directly interacting with the evidence retrieval component

[9] https://github.com/bekou/evidence_aware_nlp4if

[10] https://huggingface.co/UBC-NLP/MARBERTv2

[11] https://huggingface.co/bert-base-multilingual-uncased

[12] The model proposed recently by DeHaven and Scott (2023) is SOTA for claim verification for FEVER, but they adopted DeBERTa V2 XL MNLI, which is not available for Arabic. Moreover, we could not adopt their retrieval model due to the re-retrieval step which is not applicable to our task.

Table 2: Performance of Cross-lingual Zero-shot Evidence Retrieval. Bold scores are the best for each test set. Standard and FEVER scores to evaluate the user and system scenario respectively.

Test Set	Retrieval Model	Standard MAP	Standard R@5	FEVER P@5	FEVER R@5	FEVER F_1@5
AuRED	MLA	**0.521**	**0.589**	**0.289**	**0.755**	**0.413**
	KGAT	0.434	0.512	0.244	0.714	0.359
AuRED*	MLA	**0.619**	**0.698**	**0.266**	**0.840**	**0.401**
	KGAT	0.508	0.620	0.230	0.798	0.356

Table 3: Performance of Cross-lingual Zero-shot Rumor Verification. Bold scores are the best for each test set.

Test Set	Verification model	m-F_1	Strict m-F_1
AuRED	MLA	0.215	0.171
	KGAT	**0.422**	**0.413**
AuRED*	MLA	0.226	0.196
	KGAT	**0.426**	**0.417**

to get evidence that can help her verify a given rumor. In such scenario the system should retrieve as much evidence, preferably from different authorities, as possible to convince the user. Therefore, the system is required to provide a *ranked list* of potentially-evidence tweets. To measure the ability of the system to retrieve evidence tweets higher in the list, we adopt the *standard* information retrieval rank-based measure Mean Average Precision (MAP), and we report Recall@5 (R@5).

- The **System Scenario** is the case where the output of the retrieval component is used automatically by the down-stream rumor verification component. In this scenario, retrieving at least one evidence tweet for the given rumor might be enough. Hence we consider the evaluation measures adopted by the FEVER shared task (Thorne et al., 2018b), namely Macro R@5, where an instance is scored if at least one evidence is retrieved, and we report Macro P@5, and F_1@5 computed using both these metrics.

6.3.2 Rumor Verification

To evaluate the performance of rumor verification models, we adopt Macro-F_1 measure to account for the label imbalance in our data. Inspired by FEVER score which adopts strict label accuracy (Thorne et al., 2018b), we also adopt *strict* Macro-F_1, where we consider the label correct only if at least one correct evidence is retrieved by the adopted evidence retrieval model. Specifically, we consider an instance a *false positive* if the label is predicted correctly but no single correct evidence was retrieved.

7 Results and Discussion

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

7.1 Cross-lingual Zero-shot Scenario (RQ1)

For this setup, we fine-tuned MLA and KGAT models presented in Section 6.1 and Section 6.2 using the authors' setup for both evidence retrieval and claim verification tasks. Since, for this scenario, we train on English data (FEVER) and test on Arabic data (AuRED), we adopted mBERT as the pre-trained model. The models were then used to retrieve evidence for AuRED test rumors and verify them using the retrieved evidence. We report the average performance, using cross-validation, for evidence retrieval and rumor verification in Table 2 and Table 3 respectively.

Evidence Retrieval: As shown in Table 2, MLA achieved better performance than KGAT for evidence retrieval across all evaluation measures on both AuRED and AuRED*. Given that this setup is both cross-lingual (training and testing on two different languages -English vs. Arabic-) and cross-domain (training and testing on two different domains -Web pages vs. tweets-), we believe the performance is acceptable. It also indicates the potential of knowledge transfer using FEVER dataset to our task for evidence retrieval. Looking at the recall performance, we also note that MLA was able to retrieve about an average of 59% of the evidence tweets over all rumors, and at least one evidence tweet for about 76% of them. The latter in particular is important for the system scenario, where the evidence is used in the verification down-stream task. Overall, the models performed better on AuRED* than AuRED in terms of MAP and recall. This is somewhat expected as AuRED* is less challenging because evidence is retrieved from the timeline of a single authority for each rumor.

Rumor Verification: As presented in Table 3, the performance of both models is considered poor, which we speculate due to the domain difference. We believe the way authorities refute or support rumors in their tweets differs significantly in terms of writing from how Wikipedia sentences refute or support claims (refer to Table 7 in Appendix B.1). Recall that FEVER claims are generated by manipulating the Wikipedia sentences adopting paraphrasing, negation, or entity substitution to name a few changes (Thorne et al., 2018a). Thus, the models may have learned different styles of evidence to decide whether a given a rumor REFUTES, SUPPORTS, or NOT ENOUGH INFO to verify it. Finally, we observe that KGAT significantly outperforms MLA in verification, despite the superiority of the latter in evidence retrieval, showing clearly that the retrieval and verification models are different.

7.2 In-domain Fine-tuning Scenario (RQ2)

For this setup, we tested the evidence retrieval and rumor verification models presented in Section 6.1 and Section 6.2 respectively. We fine-tuned all the models using AuRED. The performance on evidence retrieval and rumor verification is presented in Table 4 and Table 5 respectively.

Evidence Retrieval: MLA and STAuRED are the best performing models in terms of the standard MAP and R@5 measures on both AuRED and AuRED*. The performance of STAuRED in particular highlights the potential of detecting the stance for evidence retrieval. However, surprisingly, BM25 (the lexical retrieval model) is the best performing model in retrieving evidence for *more rumors*, as indicated by the FEVER scores, on both AuRED and AuRED*. Recall that FEVER measures reward models that cover *more rumors* (by retrieving at least one evidence) higher than models that retrieve *more evidence*. This result indicates that lexical retrieval is probably enough to provide minimum evidence, however that might not be sufficient for fact-checkers who are interested in more evidence to reach a solid verification decision.

Rumor Verification: Neither of the models perform well on this task, indicating a huge room for improvement. One of the main reasons is the small number of training rumors in AuRED; in fact, only 96 rumors (constituting 3 folds) were used for training. There are multiple solutions to address this problem in the future including *data augmentation*, e.g., using synthetic data that is automatically generated by large language models (Ubani et al., 2023) or seq2seq text generation models (Pan et al., 2023), or *domain adaptation* (Yue et al., 2023) over fact checking datasets. While KGAT still exhibits better performance than MLA when fine-tuned with the in-domain training data, the performance interestingly has not reached the performance under the cross-lingual setup shown in Table 3. This can be attributed to the size of the training data in both cases; the big collection of claims in FEVER (145,449 training claims) enabled KGAT to better learn reasoning for the verification task. We will leave the investigation of such result to future work.

Table 4: Performance of In-domain Fine-tuning for Evidence Retrieval. Bold and underlined scores are the best and second-best respectively for each test set. Standard and FEVER scores to evaluate the user and system scenario respectively.

Test Set	Retrieval Model	Standard MAP	Standard R@5	FEVER P@5	FEVER R@5	FEVER F_1@5
AuRED	BM25	0.578	0.655	**0.325**	**0.892**	**0.476**
	mContriever	0.555	0.590	0.290	0.766	0.420
	MLA	**0.651**	<u>0.697</u>	<u>0.323</u>	<u>0.873</u>	<u>0.468</u>
	KGAT	0.608	0.650	0.292	0.808	0.426
	TML	0.540	0.596	0.259	0.757	0.384
	STAuRED	<u>0.622</u>	**0.700**	0.295	0.841	0.435
AuRED*	BM25	0.648	0.745	**0.326**	**0.903**	**0.479**
	mContriever	0.626	0.693	0.274	0.830	0.412
	MLA	<u>0.706</u>	<u>0.747</u>	<u>0.292</u>	0.883	<u>0.437</u>
	KGAT	0.681	0.726	0.268	0.873	0.409
	TML	0.641	0.723	0.264	<u>0.884</u>	0.407
	STAuRED	**0.715**	**0.770**	0.286	0.883	0.431

Table 5: Performance of In-domain Fine-tuning for Rumor Verification. Bold scores are the best for each test set.

Test Set	Verification model	m-$F1$	Strict m-$F1$
AuRED	MLA	0.351	0.324
	KGAT	**0.371**	**0.342**
AuRED*	MLA	0.354	0.339
	KGAT	**0.366**	**0.348**

8 Conclusion and Future Work

In this paper, we introduced the new task of *rumor verification using evidence from authorities over Twitter*. We constructed and released the first Authority-Rumor-Evidence Dataset (AuRED) which consists of 160 rumors expressed in tweets and 692 timelines of authorities Twitter accounts comprising about 34k annotated tweets in total. We explore existing fact-checking models to set up the baseline systems for our two substasks namely evidence retrieval and rumor verification. Our experiments show that evidence retrieval models for fact-checking achieved competitive benchmark results even under *cross-lingual zero-shot* setup, however the performance on rumor verification is still far from enough. For future work, we plan to (1) consider the multimodality of evidence tweets to improve the evidence retrieval, (2) augment the dataset to expand the number of rumors to improve the rumor veracity prediction, (3) propose models to improve the performance achieved on both sub-tasks, and (4) construct a similar dataset in English to facilitate and encourage research on the task.

9 Limitations

Due to time and budget constraints, this work is limited in two aspects as presented below:

Data Size. The small number of rumors in our data, despite being traditionally reasonable for retrieval tasks, make it very challenging for the rumor verification task in particular. This motivates the need to build models with the ability to transfer knowledge from relevant datasets. However, because we are targeting the Arabic language this raised another limitation due to the limited Arabic resources for fact checking and evidence-based rumor verification. Moreover, we believe data augmentation with real or synthetic data can improve the performance of the models.

Evidence multimodality. Although, evidence is not textual in 38.5% of the evidence tweets, we did not consider the multimodality in this work. Considering multimodal evidence retrieval models (Hu et al., 2023; Yao et al., 2023) or expanding the context of the rumor with extracted text from images, videos, or external news articles embedded in the authority tweets can further improve the retrieval of evidence tweets.

Acknowledgments

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

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, 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 (Volume 1: Long Papers)*, pages 7088–7105.

Maha Al-Yahya, Hend Al-Khalifa, Heyam Al-Baity, Duaa AlSaeed, and Amr Essam. 2021. Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches. *Complexity*, 2021.

Rasha M Albalawi, Amani T Jamal, Alaa O Khadidos, and Areej M Alhothali. 2023. Multimodal Arabic Rumors Detection. *IEEE Access*.

Sarah Alqurashi, Btool Hamoui, Abdulaziz Alashaikh, Ahmad Alhindi, and Eisa Alanazi. 2021. Eating Garlic Prevents COVID-19 Infection: Detecting Misinformation on the Arabic Content of Twitter. *arXiv preprint arXiv:2101.05626*.

Thayer Alshaabi, David Rushing Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. 2021. The Growing Amplification of Social Media: Measuring Temporal and Social Contagion Dynamics for over 150 Languages on Twitter for 2009–2020. *EPJ data science*, 10(1):15.

Lama Alsudias and Paul Rayson. 2020. COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Saud Althabiti, Mohammad Ammar Alsalka, and Eric Atwell. 2022. Detecting Arabic Fake News on Social Media using Sarcasm and Hate Speech in Comments.

Mohamed Seghir Hadj Ameur and Hassina Aliane. 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset. *Procedia Computer Science*, 189:232–241.

Na Bai, Fanrong Meng, Xiaobin Rui, and Zhixiao Wang. 2022a. A multi-task attention tree neural net for stance classification and rumor veracity detection. *Applied Intelligence*, pages 1–11.

Na Bai, Fanrong Meng, Xiaobin Rui, and Zhixiao Wang. 2022b. Rumor detection based on a Source-Replies conversation Tree Convolutional Neural Net. *Computing*, 104(5):1155–1171.

Na Bai, Fanrong Meng, Xiaobin Rui, and Zhixiao Wang. 2023. A multi-task attention tree neural net for stance classification and rumor veracity detection. *Applied Intelligence*, 53(9):10715–10725.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. Understanding the impact of evidence-aware sentence selection for fact checking. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 23–28, Online. Association for Computational Linguistics.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor Detection on Social Media with Bi-Directional Graph

Convolutional Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 549–556.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. Loren: Logic-regularized reasoning for interpretable fact verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10482–10491.

Jiho Choi, Taewook Ko, Younhyuk Choi, Hyungho Byun, and Chong-kwon Kim. 2021. Dynamic graph convolutional networks with attention mechanism for rumor detection on social media. *Plos one*, 16(8):e0256039.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Mitchell DeHaven and Stephen Scott. 2023. BEVERS: A general, simple, and performant framework for automatic fact verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. 2022. PHEMEPlus: Enriching social media rumour verification with external evidence. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 49–58, Dublin, Ireland. Association for Computational Linguistics.

Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2021. COVID-19-FAKES: A Twitter (Arabic/English) dataset for detecting misleading information on COVID-19. In *Advances in Intelligent Networking and Collaborative Systems: The 12th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2020) 12*, pages 256–268. Springer.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Fatima Haouari and Tamer Elsayed. 2023. Detecting Stance of Authorities Towards Rumors in Arabic Tweets: A Preliminary Study. In *Advances in Information Retrieval*, pages 430–438, Cham. Springer Nature Switzerland.

Fatima Haouari and Tamer Elsayed. 2024. Are authorities denying or supporting? Detecting stance of authorities towards rumors in Twitter. *Social Network Analysis and Mining*, 14(1):34.

Fatima Haouari, Tamer Elsayed, and Watheq Mansour. 2023. Who can verify this? Finding authorities for rumor verification in Twitter. *Information Processing & Management*, 60(4):103366.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 72–81, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Maram Hasanain, Fatima Haouari, Reem Suwaileh, Zien Sheikh Ali, Bayan Hamdan, Tamer Elsayed, Alberto Barrón-Cedeno, Giovanni Da San Martino, and Preslav Nakov. 2020. Overview of CheckThat! 2020 Arabic: Automatic Identification and Verification of Claims in Social Media. In *CLEF*.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 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*, SIGIR '23, page 2901–2912, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring Listwise Evidence Reasoning with T5 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 (Volume 2: Short Papers)*, pages 402–410.

K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2447–2460, Online. Association for Computational Linguistics.

Sumeet Kumar and Kathleen Carley. 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*, pages 5047–5058, Florence, Italy. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221, Turku, Finland. Linköping University Electronic Press.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3818–3824.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Annual Meeting of the Association for Computational Linguistics*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Zhiyuan Ma, Jianjun Li, Guohui Li, and Yongjing Cheng. 2022. GLAF: Global-to-local aggregation and fission network for semantic level fact verification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1801–1812, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ahmed Redha Mahlous and Ali Al-Laith. 2021. Fake News Detection in Arabic Tweets during the COVID-19 Pandemic. *International Journal of Advanced Computer Science and Applications*, 12(6).

P Nakov, D Corney, M Hasanain, F Alam, T Elsayed, A Barron-Cedeno, P Papotti, S Shaar, G Da San Martino, et al. 2021. Automated fact-checking for assisting human fact-checkers. In *IJCAI*, pages 4551–4558. International Joint Conferences on Artificial Intelligence.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019b. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2553–2566.

Dan S Nielsen and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023. Investigating zero-and few-shot generalization in fact verification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–524.

Eunhwan Park, Jong-Hyeon Lee, DongHyeon Jeon, Seonhoon Kim, Inho Kang, and Seung-Hoon Na. 2022. SISER: Semantic-infused selective graph reasoning for fact verification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1367–1378, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aktham Sawan, Thaer Thaher, and Noor Abu-el rub. 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)*, pages 1–6.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 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 (Volume 1: Long Papers)*, pages 1612–1622, Online. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for Evidence Retrieval and Claim Verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 359–366. Springer.

Chenguang Song, Kai Shu, and Bin Wu. 2021. Temporally evolving graph neural network for fake news detection. *Information Processing & Management*, 58(6):102712.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT. *arXiv preprint arXiv:2304.14334*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 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)*, pages 1392–1401, Online. Association for Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239, Toronto, Canada. Association for Computational Linguistics.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2019. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *International Conference on Learning Representations*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

A Comparison with Other Datasets

As presented in Table. 6, we review existing rumor verification datasets in terms of the evidence adopted for rumor verification. As shown in the table, most of existing studies focus on using the propagation networks. Some of the studies relied on the rumor textual content solely without any external evidence for verification (Alsudias and Rayson, 2020; Albalawi et al., 2023; Ameur and Aliane, 2021; Elhadad et al., 2021). Recently, some studies incorporated evidence from the Web such as relevant Web articles (Hasanain et al., 2020;

Table 6: Comparison between AuRED and existing datasets for rumor verification in social media.

Dataset	# Rumors	Platform	Evidence	Language
Arabic-COVID19 (Alsudias and Rayson, 2020)	2,000	Twitter	None	Ar
Multimodal-Rumors (Albalawi et al., 2023)	4,025	Twitter	None	Ar
COVID-19-FAKES (Elhadad et al., 2021)	220,000	Twitter	None	Ar/En
PHEME (Zubiaga et al., 2016)	330	Twitter	Propagation networks	En
RumorEval17 (Derczynski et al., 2017)	325	Twitter	Propagation networks	En
RumorEval19 (Gorrell et al., 2019)	446	Twitter/Reddit	Propagation networks	En
Twitter15/16 (Ma et al., 2017)	818	Twitter	Propagation networks	En
Weibo (Ma et al., 2016)	4,664	Weibo	Propagation networks	Zh
DAST (Lillie et al., 2019)	220	Reddit	propagation networks	Da
ArCOV19-Rumors (Haouari et al., 2021)	3,584	Twitter	Propagation networks	Ar
CheckThat!2020 (Hasanain et al., 2020)	165	Twitter	Web articles	Ar
PHEMEPlus (Dougrez-Lewis et al., 2022)	1972	Twitter	Propagation networks/Web articles	En
MuMIN (Nielsen and McConville, 2022)	12,914	Twitter	Propagation networks/Metadata	Multi
MR^2 (Hu et al., 2023)	14,700	Twitter/Weibo	Propagation networks/Web articles and images	En/Zh
AuRED	160	Twitter	Authority tweets	Ar

Dougrez-Lewis et al., 2022; Hu et al., 2023) and images (Hu et al., 2023) in addition to social media users' metadata (Nielsen and McConville, 2022). Most of the existing datasets for *Arabic* Rumor verification do not incorporate any external evidence. Some notable exceptions are the data released by Haouari et al. (2021) and Hasanain et al. (2020) who incorporated the propagation networks and Web articles as external evidence respectively. Compared to existing datasets, AuRED incorporates evidence from authority timelines.

B Data Overview

In this section, we present some examples from our dataset (B.1), discuss our data annotation challenges (B.2), show some analysis about our dataset (B.3), and finally we present the resources we release (B.4).

B.1 Data Examples

We present example rumors and corresponding evidence tweets from our AuRED dataset in Table 7.

B.2 Annotation Challenges

There are several challenges associated with annotating the data. We elaborate on a few of them through discussing the rumor tweet "Urgently, giving the Corona vaccine has stopped urgently in the Kingdom of Saudi Arabia. There is no power or strength from God. Five people died after receiving the vaccine."

Multiple rumors: A tweet may contain multiple potential rumors. For example, our tweet contains two potential rumors as a result of receiving the new Corona virus vaccine: (a) "vaccine has stopped urgently in the Kingdom of Saudi", and (b) "Five people died after receiving the vaccine". We asked annotators to focus on the rumor that had been already fact-checked by our sources (e.g., rumor (b) is verified by "Misbar" fact checking platform, assuming those are viral, consequently could have higher impact on the community.

Table 7: Sample rumors and corresponding evidence tweets (translated to English) from AuRED. The refuted and supported rumors have more than one evidence, but only one is presented for demonstration purposes. The authorities Twitter accounts, and the tweets posting dates are highlighted in green and yellow respectively.

Refuted Rumor: Moroccan reports: Bakary Gassama, is the referee of the return match between Al-Ahly and Wydad #195Sports [URL] [21-10-2020] **Authority Evidence:** [@AlAhlyTV] Learn about the biography of referee Gomez, referee of the Al-Ahly and Wydad match today YouTube: [URL] #Six #Africa_Ahly #Alahlytv [23-10-2020] **Authority Non-Evidence:** [@caf_online_AR] An exciting semi-final between Al-Ahly and Wydad Watch the four goals in a summary of the highlights of the entire match [24-10-2020]
Supported Rumor: The Libyan Ministry of Foreign Affairs' Twitter account has been hacked [URL] [22-12-2022] **Authority Evidence:** [@Mofa_Libya] The account has been officially restored. We thank everyone who contributed and cooperated with us. @GovernmentLY @Hakomitna [21-12-2022] **Authority Non-Evidence:** [@Mofa_Libya] Congratulations to the State of #Libya on the occasion of the Independence Day [24-12-2022]
Unverifiable Rumor: Watch.. how #Qataris_celebrated in the streets of Doha after the Kingdom of Saudi Arabia agreed to open the land and air borders with their country [URL] @marsdnews24 [05-01-2021] **Authority Non-Evidence:** [MBA_AlThani_] The Kuwaiti Foreign Minister announces that an agreement has been reached under which the airspace and land and sea borders between the Kingdom of Saudi Arabia and the State of Qatar will be opened as of this evening [04-01-2021]

(a) Image Evidence (b) Video Evidence (c) Implicit Evidence

Figure 3: Multimodality of evidences in AuRED.

Time sensitive rumors: The factuality of some rumors may change within a short period of time. For example, the COVID tolls (e.g., deaths) in our example could increase or decrease over time if the rumor is true, hence, we urged the annotators to consider the tweet timestamp while annotating.

Context of evidences: Verifying rumors requires looking at the authority timelines entirely rather than reading tweets independently. For instance, verifying the number of COVID tolls could require summing up the number of cases in an authority timeline within a time window.

Multimodality of evidences: Evidence could be extracted from text, images, videos, or a combination of these. The Saudi Ministry of Health posted several tweets that are useful for verification but not all of them contain textual evidences. Figure 3a shows an image of the highlights of the press conference of spokesman of the Saudi Ministry of Health. The spokesman announced the beginning of the vaccine campaign and encouraged people to register to take the vaccine which denies both rumor (a) and (b). On the other hand, Figure 3b shows a video of the health minister confirming the safety of the vaccine and denying the rumors about its side effects. The tweet also contains an implicit textual evidence that calms the public down. Accordingly, we asked annotators to carefully analyze the media not only the tweet text which required extra time and effort.

Implicit evidences: The evidences in authority timelines are not always stated explicitly. For example, Figure 3c shows a tweet from the Saudi Ministry of Health encouraging people to book vaccination appointments. Without an explicit statement, this tweet denies both rumor (a) and (b). We highly urged the annotators to consider all potential evidences including implicit ones.

B.3 Data Analysis

To show the quality of AuRED, we analyzed its coverage and diversity to ensure the generalizability of models trained on it. In the following we discuss different aspects.

Dialectical/Geographical Coverage: AuRED contains rumors that are of interest to different Arab countries such as Egypt, Qatar, Saudi Arabia, Kuwait, among other countries. Figure 4 shows the geographical distribution of rumors across the Arab countries. The dataset also covers rumors of interest to the Arab users although not happening in the Arab region. Such geographical coverage implies the coverage of diverse dialects in AuRED. We used ASAD tool (Hassan et al., 2021) to automatically analyze the dialectical coverage of the tweets in AuRED. We found 92.5% of tweets are written in Modern Standard Arabic (MSA) and the remaining are dialectical tweets.

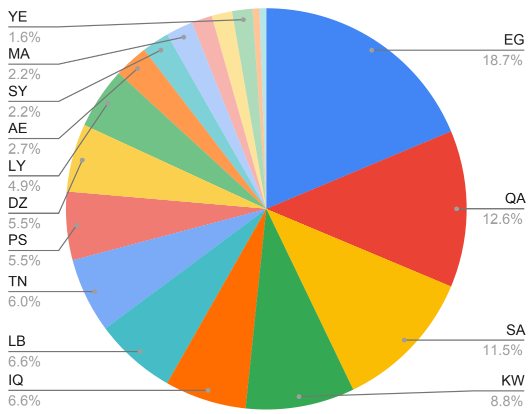

Figure 4: Geographical coverage of rumors in AuRED. The countries are represented by their 2-letter ISO codes.

Domain Coverage: We define *domain* here as the topic of the rumor such as politics, health, sports,

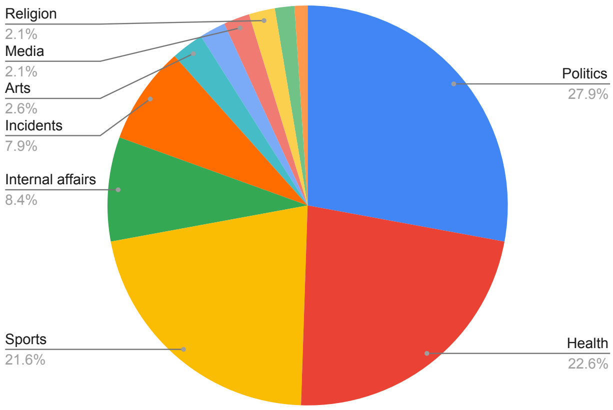

Figure 5: Domain Coverage of rumors in AuRED.

etc. Figure 5 shows the diverse coverage of domains of rumors in AuRED.

Multimodality: To support the development of versatile verification systems, AuRED is labeled for different types of evidences, i.e., text, and or media. It contains 49.05% multimodal evidence tweets, 38.5% of which are media evidences that show the insufficiency of text for rumor verification. The remaining contain both text and media that complement each other for rumor verification.

B.4 Data Release

We release the following data as part of AuRED, taking into consideration the content distribution policy:[13]

- **Rumors:** 160 rumors expressed in tweets each labeled as SUPPORTES, REFUTES, or NOT ENOUGH INFO. We release the rumor IDs and tweets text.
- **Authorities timelines:** Each rumor is associated with timelines of potential authorities. We release the authority Twitter account link, tweets IDs, and tweets text.
- **Evidence tweets:** Each rumor is associated with the evidence tweets IDs and text.
- **Authorities tweets media:** The images and videos extracted from authority tweets.
- **Data folds:** To enable consistent benchmarking on the dataset, we provide our data folds. i.e., 5 folds we adopted for our cross-validation setup.
- **Annotation guidelines:** We share our language-independent evidence retrieval annotation guidelines to encourage the construction of similar collections in other languages.
- **Benchmarks Code**: For reproducibility and to facilitate research on the task we release our source code.

[13] https://developer.twitter.com/en/developer-terms/agreement-and-policy