Overview of the CLEF-2024 CheckThat! Lab Task 5 on Rumor Verification using Evidence from Authorities

Notebook for the CheckThat! Lab at CLEF 2024

Fatima Haouari[1], Tamer Elsayed[1] and Reem Suwaileh[2]

[1] *Qatar University, Doha, Qatar*
[2] *Hamad Bin Khalifa University, Doha, Qatar*

Abstract

We present an overview of Task 5 of the seventh edition of the CheckThat! Lab, which is a part of the 2024 Conference and Labs of the Evaluation Forum (CLEF). In the Rumor Verification using Evidence from Authorities task, given a rumor expressed in a tweet and a set of authorities Twitter accounts for that rumor, participating systems should retrieve up to 5 evidence tweets posted by those authorities, and determine the veracity of the rumor according to the retrieved evidence. A total of 3 and 5 teams submitted their runs (5 and 11 runs) for Arabic and for English, respectively, out of which 2 made submissions for both languages. In this paper, we present our data construction approach, evaluation setup, and an overview of the participating systems. We publicly release all the datasets and evaluation scripts to promote further research on this task.

Keywords

Fact Checking, Claims, Social Media

1. Introduction

The CheckThat! lab runs for the seventh time under the umbrella of CLEF 2024 [1, 2]. In this edition of the lab six tasks were offered: task 1 on Check-Worthiness Estimation, task 2 on subjectivity detection, task 3 on persuasion techniques, task 4 on detecting hero, villain, and victim from memes, task 5 on rumor verification using evidence from authorities (this paper), and task 6 on robustness of credibility assessment with adversarial examples (InCrediblAE).

In this paper, we describe in detail Task 5 of this year's lab,[1] *Rumor Verification using Evidence from Authorities.* Task 5 is defined as follows: "*Given a rumor expressed in a tweet and a set of authorities (one or more authority Twitter accounts) for that rumor, represented by a list of tweets from their timelines during the period surrounding the rumor, the system should retrieve up to 5 evidence tweets from those timelines, and determine if the rumor is supported (true), refuted (false), or unverifiable (in case not enough evidence to verify it exists in the given tweets) according to the evidence.*"

The rest of this paper is organized as follows. We give an overview of the related work in Section 2. We define our task, and present our adopted evaluation measures in Section 3, and Section 4 respectively. We present a full overview about the Arabic shared task and the English shared task in Section 5 and Section 6 respectively, including our datasets construction approach, an overview of the participants' systems, and discuss the evaluation results . Finally, we conclude in Section 7.

2. Related Work

A large number of existing studies in the broader literature have studied rumor verification in social media [3, 4, 5, 6, 7, 8, 9]. Most early studies has incorporated the propagation networks such as the structure of replies [6, 10, 7, 11, 8], stance of replies [12, 13, 3, 4, 5], or retweeters metadata [9] as a

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.
200159617@qu.edu.qa (F. Haouari); telsayed@qu.edu.qa (T. Elsayed); rsuwaileh@hbku.edu.qa (R. Suwaileh)

[1] Refer to [2] for an overview of the full CheckThat! 2024 lab.

source of evidence. Some authors have also suggested that evidence from the Web [14, 15], or stance of authority tweets towards rumors [16, 17] can further improve the automatic rumor verification. Rumor verification in social media was addressed in multiple languages mainly in English [12, 14, 4, 3] or Chinese [18, 6, 7]. However, *Arabic* rumor verification is still under-studied. Most of the existing studies relied on the rumor textual content solely for verification [19, 20, 21, 22]. Recently, Haouari et al. [11] exploited the replies structure, Althabiti et al. [23] proposed detecting sarcasm and hate speech in the replies, while Albalawi et al. [24] leveraged the images and videos embedded in the rumor tweet. Differently, we propose incorporating the evidence tweets retrieved from the authority timelines for *Arabic* and *English* rumor verification.

3. Task Definition

The *Rumor Verification using Evidence from Authorities* task consists of two subtasks defined as follows:

- **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.
- **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. Evaluation Measures

Evidence retrieval. The official evaluation measure for evidence retrieval is Mean Average Precision (MAP). We also report Recall@5.

Rumor Verification. We use the Macro-F1 to evaluate the classification of the rumors. Additionally, we consider a Strict Macro-F1 where the rumor label is considered correct only if at least one retrieved authority evidence is correct.

5. Arabic Shared Task

In this section, we give an overview about the *Arabic* shared task. We present our dataset construction approach in Section 5.1. The approaches adopted by the participating systems, and their evaluation results are presented in Section 5.2, and Section 5.3 respectively.

5.1. Dataset

To construct our dataset,[2] we randomly selected 160 rumors from two existing datasets namely AuFIN [25, 26] and AuSTR [17]. Specifically, we selected 99 (61.9%) from AuFIN and 61 (38.1%) from AuSTR. We then annotated the dataset following two steps 1) finding authorities that may tweet evidence that can help in rumor verification (Section 5.1.1), and 2) evidence extraction including the authority timelines collection and annotation (Section 5.1.2). Our task dataset, covers 160 rumors annotated with their corresponding 692 authority timelines, comprising about 34k annotated tweets in total. We randomly split the data into 96 training, 32 development, and 32 test examples.

[2] https://gitlab.com/checkthat_lab/clef2024-checkthat-lab/-/tree/main/task5

5.1.1. Authority Finding

The task is proposed recently by Haouari et al. [25]. 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 Sports event in Qatar, then ministry and minister of Sports and Youth, and managers of the event 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. [25]. Two annotators, co-organizers of this task, performed the task independently, then met to discuss their annotations. Only potential authorities that both annotators agreed upon during their meeting were kept in our data.

5.1.2. Evidence extraction

To collect the authority timelines, we used the Twitter Academic search API which facilitates collecting users historical timelines.[3] We consider the rumor tweet as a pointer to the time span of the rumor propagation, where we assume that the rumor is circulating for a few days before and/or after the rumor tweet posting time. Therefore, we limit the authority timelines to the tweets within 3 days before and after the rumor tweet posting time. To extract the evidence from the collected timelines, we performed two steps:
(1) 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. 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. At the end of this stage, we measured the data quality of the double-annotated sample using Cohen's Kappa for inter-annotator agreement [27] as 0.67, which indicates "substantial" agreement [28]. It is worth mentioning, that any disagreement between the annotators was then resolved in the next stage.
(2) Resolving Disagreements: As a final step, both annotators met to discuss and resolve any disagreements in the double-annotated sample, and hence decide the final labels. Refer to [29] for more details about our data construction process.

5.2. Overview of the Participating Systems

For the *Arabic* shared task, 3 teams submitted a total of 5 runs. In the following, we present their proposed approaches.

bigIR. This team submitted 2 runs adopting two SOTA models for fact checking namely MLA [30] and KGAT [31]. For evidence retrieval, MLA is a BERT-based binary classifier fine-tuned to classify whether an authority tweet is an evidence or non-evidence, and the input to the model are pairs of (rumor_tweet, authority_tweet). At training they considered only a sample of non-evidence tweets for each rumor.[4] At inference, every (rumor_tweet, authority_tweet) of the test set is passed to the fine-tuned model and softmax scores were used to get the top N authority tweets. KGAT model is also BERT-based model, however the margin ranking loss is adopted to maximize the distance between the positive and the negative (rumor_tweet, authority_tweet) pairs. At inference, the scores obtained by passing each pair is used to rank the authority tweets.

The top 5 retrieved evidence tweets are then used to fine-tune their rumor verification model, adopting the MLA and KGAT claim verification models. KGAT is reasoning model adopting Kernel Graph Attention Network to construct a fully connected graph using the retrieved evidence. MLA on the other hand, adopts multi-task learning considering the verification as the main task, and evidence

[3] https://developer.x.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all
[4] They set the number of negative examples as 4 times the number of positive examples.

retrieval as an auxiliary task. For all their evidence retrieval and rumor verification models they used MARBERTv2 [32],[5] an Arabic BERT model pre-trained using 1 billion Arabic tweets. They fine-tuned all models for 5 epochs using a batch of size 8 using 4 different learning rates [2e-5, 3e-5, 4e-5, 5e-5]. They selected the best evidence retrieval and rumor verification models on the dev set based on Mean Average Precision (MAP) and Macro F1 respectively.

IAI Group. This group submitted 2 runs, adopting a zero-shot setup for both evidence retrieval and rumor verification. For evidence retrieval, they experimented with two approaches namely 1) using ColBERT-XM [33], a multilingual pre-trained model for semantic search, and 2) using cross-encoders. For rumor verification, in both runs, they leveraged the xlm-roberta-nli which is a RoBERTa model pre-trained with a combination of Natural Language Inference (NLI) data in multiple languages [34].

SCUoL. This team submitted 1 run where they focused solely on the rumor verification subtask. They leveraged an existing Arabic fact checking system [35], where they passed the rumor tweet solely to the system to get the veracity label.

5.3. Evaluation Results

In this section, we present and discuss the results of the participating systems for both evidence retrieval and rumor verification against our baseline.

Baseline: We adopted KGAT [31], a SOTA model for fact-checking. We fine-tuned both its evidence retrieval and rumor verification models on FEVER English fact-checking dataset [36] following the authors setup but using multilingual BERT (mBERT) [37].[6] We then test on our *Arabic* test data.

Evidence Retrieval: As presented in Table. 1, 4 out 5 runs managed to outperform the baseline significantly. bigIR team primary run outperformed all models in terms of all evaluation measures, fine-tuning their model on AuRED data. We can also observe that although IAI Group adopted a zero-shot approach significantly outperformed the baseline. Moreover bigIR secondary run which is the model used as a baseline. i.e., KGAT, but fine-tuned on AuRED show a big improvement which shows the importance of in-domain data for the task.

Table 1
Evidence retrieval (**Arabic**) official evaluation results, in terms of MAP, and Recall@5. The teams are ranked **only** based on their primary runs by the official evaluation measure MAP. Submissions with a + sign indicate submissions by task organisers.

Rank	Team (run ID)	MAP	Recall@5
1	bigIR+ (bigIR-MLA-Ar)	0.618	0.673
-	IAI Group (IAI-Arabic-Crossencoder)	0.586	0.601
2	IAI Group (IAI-Arabic-COLBERT)	0.564	0.581
-	bigIR+ (bigIR-KGAT-Ar)	0.560	0.625
	Baseline	0.345	0.423
3	SCUoL (SCUoL)	-	-

Rumor Verification: As presented in Table. 2, we observe that IAI Group primary and secondary runs outperformed all runs significantly although adopting a zero-shot approach. The results highlight that even the two models fine-tuned on the task data, bigIR models, could not achieve comparable

[5] https://huggingface.co/UBC-NLP/MARBERTv2
[6] https://huggingface.co/bert-base-multilingual-uncased

results to the best performing model. We observe that one of the bigIR models outperforms the baseline on Macro F1 only but could not beat it in terms Strict Macro F1, while the second could not even beat the baseline across all measures. This could be attributed to the small number of training examples. i.e., 96 rumors only. Finally, we observe that the run submitted by SCUol team achieved better than the baseline although not considering the authority evidence.

Table 2
Rumor verification (**Arabic**) official evaluation results, in terms of Macro F1, and Strict Macro F1. The teams are ranked **only** based on their primary runs by the official evaluation measure Macro F1. Submissions with a + sign indicate submissions by task organisers.

Rank	Team (run ID)	Macro F1	Strict Macro F1
1	IAI Group (IAI-Arabic-COLBERT)	0.600	0.581
-	IAI Group (IAI-Arabic-Crossencoder)	0.460	0.433
2	bigIR$^+$ (bigIR-MLA-Ar)	0.368	0.300
3	SCUoL (SCUoL)	0.355	-
	Baseline	0.347	0.347
-	bigIR$^+$ (bigIR-KGAT-Ar)	0.258	0.258

6. English Shared Task

This year, a major extension for the Authority Finding task is running over English data in addition to Arabic. As English is a globally dominant language, this attracted more researchers, developers, and participants, thereby increasing the task's visibility and impact. In this section, we present our dataset construction approach (Section 6.1), participating systems (Section 6.2), and evaluation results (Section 6.3) of the English Shared Task.

6.1. Dataset

To construct the English dataset, we translated the Arabic dataset (refer to Section 5.1). The rationale behind this approach is that topics and issues concerning the Arab region are frequently discussed in English, especially by Arab users who communicate in English on Twitter. Additionally, international journalists who do not speak Arabic are interested in ongoing discussions in the region. Therefore, translations help us capture representative rumors that can be discussed within English content on Twitter while also reducing the annotation effort. We have followed a two stage process of automatic translation and manual validation that we discuss in the following.

Automatic translation We automatically translated the entire Arabic dataset using Googletrans library.[7] We translated all 160 rumor tweets, and their associated authority tweets.

Manual Validation While automatic translation can expedite the development of monolingual and cross-lingual authority finding systems, it introduces several challenges that could affect the quality and reliability of the resulting data. To address this, we manually validated the translations of a random sample of tweets. Specifically, we reviewed the translations of all rumors and a sample of 2,138 tweets from authorities timelines. We edited 514 (24%) tweets to correct errors and inaccuracies while 1,624 tweets (75.96%) remained unedited.

[7] https://py-googletrans.readthedocs.io/en/latest/

Challenges Through the validation process, we have observed issues and challenges that we addressed to maintain the quality and reliability of the dataset. We discuss a few in the following:

- *Inaccuracies*: Automatic translation tools can produce inaccurate translations, especially for complex sentences, idiomatic expressions, and context-dependent phrases. These inaccuracies can lead to errors in the dataset.
- *Loss of Nuance and Context*: Automatic translations may fail to capture the nuanced meanings and cultural context of the original text. This can result in a loss of important information.
- *Inconsistencies*: Automatic translations may change for the same text, leading to inconsistencies within the same dataset.

Despite these challenges, we opted to enable the development of English systems and consider better approaches for constructing the dataset in the future.

6.2. Overview of the Participating Systems

For the *English* shared task, 5 teams submitted a total of 11 runs. In the following, we present their proposed approaches.

AuthEv-LKolb [38]. This team participated with 3 runs. They adopted OpenAI GPT-4 assistant for rumor verification in all their runs, where they pass each single rumor-evidence pair prompting GPT-4 to return a judgement and a confidence. The N judgements are then combined into a final label. For evidence retrieval for two of their runs, they adopted OpenAI embeddings and computed the cosine similarity between the embedding vectors of the rumor tweet and each authority tweet to get the closest top N. In their third run, they adopted a simple PyTerrier BatchRetrieve pipeline of BM25 and PL2 to retrieve the top evidence tweets. The authors in one of their runs used external data where they collected the authorities Twitter account information, and augmented the input text with the authority name and bio for both the evidence retrieval and rumor verification.

Axolotl [39]. They submitted 3 runs, where they adopted for evidence retrieval BM25 lexical retrieval. To retrieve for relevant authority tweets, they give importance to hashtags in the rumor tweet by boosting them with respect to just text. For two of their runs they further reranked the top retrieved tweets using either sentence-t5-base,[8] or Llama3 8B. For rumor verification, they adopted a stance-based approach using either Llama3 8B or all-mpnet-base-v2.[9]

bigIR. The team participated with 2 runs adopting the same models and setup used for Arabic. However, they replaced MARERTv2 with the English BERT base [37].[10]

DEFAULT [40]. They formulated the task as a retrieval-augmented classification and jointly trained the rumor verification classifier and the evidence retriever. They fine-tuned ColBERT [41] and used MaxSim score as the similarity score.

IAI Group. They adopted a zero-shot setup for both their runs. For evidence retrieval they adopted either ColBERT or cross-encoders. For rumor verification they exploited a RoBERTa model pre-trained with NLI task data.

6.3. Evaluation Results

In this section, we present and discuss the results of the participating systems for both evidence retrieval and rumor verification against our baseline.

[8] https://huggingface.co/sentence-transformers/sentence-t5-base
[9] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[10] https://huggingface.co/google-bert/bert-base-uncased

Table 3
Evidence retrieval (**English**) official evaluation results, in terms of MAP, and Recall@5. The teams are ranked **only** based on their primary runs by the official evaluation measure MAP. Submissions with a + sign indicate submissions by task organisers.

Rank	Team (run ID)	MAP	Recall@5
-	IAI Group (IAI-English-Crossencoder)	0.628	0.676
1	bigIR+ (bigIR-MLA-En)	0.604	0.677
2	Axolotl (run_rr=llama_sp=llama_ rewrite=3_boundary=0,4_hashtagW=1)	0.566	0.617
3	DEFAULT (DEFAULT-Colbert1)	0.559	0.634
4	IAI Group (IAI-English-COLBERT)	0.557	0.590
5	AuthEv-LKolb (AuthEv-LKolb-oai)	0.549	0.587
-	bigIR+ (bigIR-KGAT-En)	0.537	0.618
-	AuthEv-LKolb (AuthEv-LKolb-terrier-oai-preprocessing)	0.524	0.563
-	AuthEv-LKolb (AuthEv-LKolb-oai-extdata)	0.510	0.619
-	Axolotl (run_rr=dl_sp=llama_ rewrite=0_boundary=0,2_hashtagW=1)	0.489	0.545
-	Axolotl (run_rr=none_sp=dl_ rewrite=0_boundary=0,1_hashtagW=1)	0.489	0.545
	Baseline	0.335	0.445

Baseline: We adopted the same model fine-tuned for the *Arabic* shared task baseline (refer to Section 5.3), but we tested on our *English* test data.

Evidence Retrieval: As shown in Table. 3, all the submitted runs outperformed our baseline. Looking at the primary runs, we observe that the models fine-tuned using the task data, bigIR-MLA-En and DEFAULT-Colbert1 runs, got the 1^{st} and 3^{rd} place respectively. The results also highlights that although Axolotl team run achieved a 2^{nd} position, bigIR outperforms it with a big margin. Interestingly, the IAI Group secondary run, under zero-shot setup, improved the retrieval in terms of MAP compared to the leading team but could not improve the recall of evidence.

Table 4
Rumor verification (**English**) official evaluation results, in terms of Macro F1, and Strict Macro F1. The teams are ranked **only** based on their primary runs by the official evaluation measure Macro F1. Submissions with a + sign indicate submissions by task organisers.

Rank	Team (run ID)	Macro F1	Strict Macro F1
-	AuthEv-LKolb (AuthEv-LKolb-oai-extdata)	0.895	0.876
1	AuthEv-LKolb (AuthEv-LKolb-oai)	0.879	0.861
-	AuthEv-LKolb (AuthEv-LKolb-terrier-oai-preprocessing)	0.831	0.831
2	Axolotl (run_rr=llama_sp=llama_rewrite=3_boundary=0,4_hashtagW=1)	0.687	0.687
-	Axolotl (run_rr=dl_sp=llama_rewrite=0_boundary=0,2_hashtagW=1)	0.630	0.570
-	Axolotl (run_rr=none_sp=dl_rewrite=0_boundary=0,1_hashtagW=1)	0.574	0.492
	Baseline	0.495	0.495
3	DEFAULT (DEFAULT-Colbert1)	0.482	0.454
-	IAI Group (IAI-English-Crossencoder)	0.459	0.444
4	bigIR+ (bigIR-MLA-En)	0.458	0.428
5	IAI Group (IAI-English-COLBERT)	0.373	0.373
-	bigIR+ (bigIR-KGAT-En)	0.373	0.373

Rumor Verification: As presented in Table. 4, only 2 teams were able to outperform the baseline, AuthEv-LKolb and Axolotl, adopting LLMs namely GPT4 or Llama respectively. The results highlight that the models adopting the fine-tuning setup (bigIR and DEFAULT models), or zero-shot setup using pre-trained language models (IAI group models) could not outperform the baseline. We can conclude that, adopting LLMs can perform well on the verification task achieving Macro F1 of 0.895. However, further investigation is required to compare their performance against models fine-tuned using the task data but with a large number of rumors.

7. Conclusion

In this paper, we presented a detailed overview of the CLEF 2024 CheckThat! Lab Task 5 for Rumor Verification using Evidence from authorities. For evidence retrieval, participants adopted either a zero-shot setup or a fine-tuning setup using the task data. For the zero-shot setup they leveraged existing pre-trained language models, LLMs, traditional lexical retrieval such BM25, or combination of these models. For rumor verification, only the models adopting LLMs managed to outperform the baseline. As a future work, we plan to enlarge the task dataset and incorporate more languages.

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 work of Reem Suwaileh is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, which is a part of Qatar Research Development and Innovation Council (QRDI). The statements made herein are solely the responsibility of the authors.

References

[1] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[2] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[3] S. Kumar, K. Carley, Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5047–5058. URL: https://aclanthology.org/P19-1498. doi:10.18653/v1/P19-1498.

[4] N. Bai, F. Meng, X. Rui, Z. Wang, A multi-task attention tree neural net for stance classification and rumor veracity detection, Applied Intelligence (2022) 1–11.

[5] S. Roy, M. Bhanu, S. Saxena, S. Dandapat, J. Chandra, gDART: Improving rumor verification in social media with Discrete Attention Representations, Information Processing & Management 59 (2022) 102927.

[6] J. Ma, W. Gao, K.-F. Wong, 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), 2018, pp. 1980–1989.

[7] J. Choi, T. Ko, Y. Choi, H. Byun, C.-k. Kim, Dynamic graph convolutional networks with attention mechanism for rumor detection on social media, Plos one 16 (2021) e0256039.

[8] N. Bai, F. Meng, X. Rui, Z. Wang, Rumor detection based on a Source-Replies conversation Tree Convolutional Neural Net, Computing 104 (2022) 1155–1171.

[9] Y. Liu, Y.-F. Wu, 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, 2018.

[10] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, J. Huang, Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 549–556.

[11] F. Haouari, M. Hasanain, R. Suwaileh, T. Elsayed, ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection, in: N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, S. Touileb (Eds.), Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 2021, pp. 72–81. URL: https://aclanthology.org/2021.wanlp-1.8.

[12] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, P. Tolmie, Analysing how people orient to and spread rumours in social media by looking at conversational threads, PloS one 11 (2016) e0150989.

[13] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 69–76.

[14] J. Dougrez-Lewis, E. Kochkina, M. Arana-Catania, M. Liakata, Y. He, PHEMEPlus: Enriching Social Media Rumour Verification with External Evidence, in: Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER), 2022, pp. 49–58.

[15] X. Hu, Z. Guo, J. Chen, L. Wen, P. S. Yu, 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, Association for Computing Machinery, New York, NY, USA, 2023, p. 2901–2912. URL: https://doi.org/10.1145/3539618.3591896. doi:10.1145/3539618.3591896.

[16] F. Haouari, T. Elsayed, Detecting Stance of Authorities Towards Rumors in Arabic Tweets: A Preliminary Study, in: Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 430–438.

[17] F. Haouari, T. Elsayed, Are authorities denying or supporting? Detecting stance of authorities towards rumors in Twitter, Social Network Analysis and Mining 14 (2024) 34.

[18] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 3818–3824.

[19] M. K. Elhadad, K. F. Li, F. Gebali, COVID-19-FAKES: A Twitter (Arabic/English) Dataset for Detecting Misleading Information on COVID-19, in: International Conference on Intelligent Networking and Collaborative Systems, Springer, 2020, pp. 256–268.

[20] A. R. Mahlous, A. Al-Laith, Fake News Detection in Arabic Tweets during the COVID-19 Pandemic, International Journal of Advanced Computer Science and Applications 12 (2021).

[21] M. Al-Yahya, H. Al-Khalifa, H. Al-Baity, D. AlSaeed, A. Essam, Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches, Complexity 2021 (2021).

[22] A. Sawan, T. Thaher, N. Abu-el rub, Sentiment Analysis Model for Fake News Identification in Arabic Tweets, in: 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT), 2021, pp. 1–6.

[23] S. Althabiti, M. A. Alsalka, E. Atwell, Detecting Arabic Fake News on Social Media using Sarcasm and Hate Speech in Comments (2022).

[24] R. M. Albalawi, A. T. Jamal, A. O. Khadidos, A. M. Alhothali, Multimodal Arabic Rumors Detection, IEEE Access (2023).

[25] F. Haouari, T. Elsayed, W. Mansour, Who can verify this? Finding authorities for rumor verification

in Twitter, Information Processing & Management 60 (2023) 103366.

[26] F. Haouari, Z. Sheikh Ali, T. Elsayed, 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, 2023.

[27] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and psychological measurement 20 (1960) 37–46.

[28] J. R. Landis, G. G. Koch, The Measurement of Observer Agreement for Categorical Data, Biometrics (1977) 159–174.

[29] F. Haouari, T. Elsayed, R. Suwaileh, AuRED: Enabling Arabic Rumor Verification using Evidence from Authorities over Twitter, in: Proceedings of ArabicNLP 2024, 2024.

[30] C. Kruengkrai, J. Yamagishi, X. Wang, A multi-level attention model for evidence-based fact checking, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2447–2460.

[31] Z. Liu, C. Xiong, M. Sun, Z. Liu, Fine-grained fact verification with kernel graph attention network, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7342–7351.

[32] M. Abdul-Mageed, A. Elmadany, et al., ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 7088–7105.

[33] A. Louis, V. Saxena, G. van Dijck, G. Spanakis, Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval, arXiv preprint arXiv:2402.15059 (2024).

[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451.

[35] S. Althabiti, M. A. Alsalka, E. Atwell, Ta'keed: The first generative fact-checking system for arabic claims, arXiv preprint arXiv:2401.14067 (2024).

[36] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, 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), 2018, pp. 809–819.

[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[38] L. Kolb, A. Hanbury, AuthEv-LKolb at CheckThat! 2024: A Two-Stage Approach To Evidence-Based Social Media Claim Verification, in: [42], 2024.

[39] A. Pasin, N. Ferro, SEUPD@CLEF: Team Axolotl on Rumor Verification using Evidence from Authorities, in: [42], 2024.

[40] S. Adhikari, H. Sharma, R. Kumari, S. Satapara, M. Desarkar, DEFAULT at CheckThat! 2024: Retrieval Augmented Classification using Differentiable Top-K Operator for Rumor Verification based on Evidence from Authorities, in: [42], 2024.

[41] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.

[42] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.