QATAR UNIVERSITY

COLLEGE OF ENGINEERING

# RETRIEVAL OF AUTHORITIES AND THEIR EVIDENCE FOR RUMOR VERIFICATION IN ARABIC SOCIAL MEDIA

BY

FATIMA HAOUARI

A Dissertation Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Doctorate of Philosophy in Computer Science

June 2024

# COMMITTEE PAGE

The members of the Committee approve the Dissertation of

Fatima Haouari defended on 29/05/2024.

Dr. Tamer Elsayed
Dissertation Supervisor

Prof. Eric Atwell
Committee Member

Prof. Rehab Duwairi
Committee Member

Prof. Cagatay Catal
Committee Member

Prof. Tamer Khattab
Committee Member

Approved:
Khalid Kamal Naji, Dean, College of Engineering

# ABSTRACT

Haouari, Fatima., Doctorate : June : 2024, Doctorate of Philosophy in Computer Science
Title: Retrieval of Authorities and their Evidence for Rumor Verification in Arabic Social Media
Supervisor of Dissertation: Dr. Tamer Elsayed.
Social media platforms have become a medium for rapidly spreading rumors along with emerging events. Those rumors may have a lasting effect on users' opinion even after it is debunked, and may continue to influence them if not replaced with convincing evidence. Journalists, or even normal users, who attempt to verify a rumor over social media, try to find a trusted source of evidence that can help them confirm or deny that specific rumor. A strong source of evidence for verifying a rumor is an authority who has the "real knowledge or power" to verify it if asked to. This dissertation contributes towards addressing the problem of rumor verification in social media. We propose augmenting the traditional rumor verification pipeline, which considers the propagation networks and the Web as sources of evidence, by incorporating authorities as another source of evidence. Specifically, in this dissertation we introduce the problem of rumor verification using evidence from authorities which we believe can help fact-checkers and automated rumor verification systems to find the right authorities and evidence from their Twitter timelines, hence helping in the verification process. First, we propose authority finding in Twitter. We then suggest incorporating those retrieved authorities by detecting their stance towards rumors in Twitter, and retrieving evidence from their timeline tweets. Finally, we propose rumor verification using evidence retrieved from those authorities. To address the problem, we construct and release three datasets targeting the Arabic language namely 1) the first Authority FINding in Twitter (AuFIN) which comprises 150 rumors (expressed in tweets) associated with a total of 1,044 authority accounts and a user collection of 395,231 Twitter accounts (members of 1,192,284 unique Twitter lists), 2) the first Authority STance towards Rumors (AuSTR) which comprises 811 (rumor tweet, authority tweet) pairs relevant to 292 unique rumors, 3) the first Authority-Rumor-Evidence Dataset (AuRED) which comprises 160 rumors expressed in tweets and 692 Twitter timelines of authorities comprising about 34k annotated tweets in total. We propose a hybrid retrieval authority finding model that combines lexical and semantic signals in addition to user profiles and network features. Furthermore, we investigate the usefulness of existing Arabic datasets for stance towards claims for detecting the stance of authorities. Finally, we study the effectiveness of existing fact-checking models for evidence retrieval from authorities and rumor verification using the retrieved evidence. Our experimental results suggest that Twitter lists and network features such as followers, and followees count, adopted previously for topic expert finding models, play a crucial role in authority finding; however, they are insufficient. This motivates the need to explore other features to differentiate experts from authorities. Moreover, our proposed hybrid model incorporating lexical, semantic, and user network features achieved a modest performance, 0.41 as precision at depth 1, which indicates that finding authorities is a challenging task, and that there is still room for continued enhancement. Our results also highlighted that adopting existing Arabic stance datasets for claim verification is somewhat useful but clearly insufficient for detecting the stance of authorities. Moreover, we found that AuSTR solely, despite the limited size, can be sufficient for detecting the stance of authorities achieving a performance of 0.84 macro-F1 and 0.78 F1 on debunking tweets. Our investigation on the effectiveness

of existing fact-checking (claim verification using evidence from Wikipedia pages) models on our problem highlighted that although evidence retrieval for fact-checking models perform relatively well on evidence retrieval from authorities, establishing strong baselines achieving 0.70 as recall at depth 5, there is still a big room for improvement. However, existing claim verification for fact-checking models perform poorly on rumor verification using evidence from authorities, 0.42 as macro-F1, no matter how good the retrieval performance is. Moreover, existing fact-checking datasets showed a potential in transfer learning to our problem, however, further investigation using different setups and datasets is required. Furthermore, drawing upon our experiments, we discuss failure factors and make recommendations for future research directions in addressing this problem. Additionally, our approach establishes a strong baseline for future studies targeting automatic rumor verification in social media, and our constructed datasets can facilitate further research on the problem. Finally, our proposed system can be integrated into verification systems, and can be also exploited by fact-checkers or journalists to find trusted sources of evidence.

# DEDICATION

*To my father who has always taught me to strive for excellence.*

# ACKNOWLEDGMENTS

First and foremost, I would like to thank the Almighty for giving me the strength and patience to work on this dissertation.

I would like to express my deepest appreciation to my supervisor Dr. Tamer Elsayed for his mentorship, constant support, invaluable patience, and for generously sharing his extensive academic knowledge and playing a pivotal role in helping me establish myself within the academic community. From the initial stages of refining my research proposal to the final submission of my dissertation, his unwavering presence and wealth of wisdom have been instrumental in shaping my academic growth. I am profoundly grateful for the immeasurable contributions he made to my development.

I am also grateful to the bigIR group members for a cherished time spent together, and being my inspiration. Their hard work and determination have motivated me to excel in my work. Heartfelt thanks to Reem Suwaileh for her kind help and treasured support during the difficult moments. Thanks should also go to Rana Malhas for her continuous moral support, and life advice.

Finally, I would like to thank my family. I could not have undertaken this journey without the unconditional love and support, and continuous prayers of my mother. I express my profound gratitude to my beloved husband Houssem who has been a source of strength, support, patience, and motivation for me throughout this entire experience. I appreciate my kids, Adel and Mariam, for their understanding when I was distracted or not fully present for them during this journey.

Finally, this dissertation was made possible by the Graduate Sponsorship Research Award (GSRA) grant# GSRA6-1-0611-19074 from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

In the last decade, microblogging platforms (e.g., Twitter[1]) have become a major source of information. As per Global Digital Report statistics[2], social media users have grown to a global total of 4.88 billion by the start of July 2023, which is almost 60.6% of the world's population. This led news agencies and traditional newspapers to move to social media in order to cope with the societal change. However, social media have also become a medium to disseminate rumors and misinformation. According to a study by Knight Foundation[3], Americans think that two-thirds of the news posted on social media are rumors. Moreover, it has been shown that rumors spread faster and deeper in social media [1], and that once users absorb rumors, it will be difficult to change their beliefs even when the rumors are resolved [2].

Rumors are defined as circulating claims whose veracity is not yet known at the time of posting [3]. A rumor is considered unverified until credible sources confirm its veracity, or provides evidence supporting it, such as eyewitnesses. Zubiaga, Liakata, Procter, *et al.* [4] found that unverified rumors spread quickly and largely at early stages. They also found that users tend to support unverified rumors during the first few minutes, and that the number of users supporting a rumor decreases after the rumor resolution. Interestingly, this includes normal users as well as news agencies accounts. Vosoughi, Roy, and Aral [1] found that false rumors significantly disseminate faster than true rumors. These findings together motivate the importance of early rumor verification in social media.

Journalists, or even normal users, who attempt to verify a rumor over social media try to find a trusted source of evidence (relevant to that rumor) that can help them confirm or deny that specific rumor. Trusted sources could be credible Web articles or Social media accounts, or even eyewitnesses. A stronger source of evidence for verifying a rumor are authorities who have the "real knowledge" to verify it if asked to. For example, FIFA is an authority for verifying a rumor about the change of date of the opening match of the Qatar 2022 World Cup. Having an automated system for finding authority accounts from Twitter for a given rumor would be a great asset in that regard. Moreover, having such a service can be extended to automatically find evidence from the accounts of those authorities. Retrieving such evidence can improve the automated veracity prediction and provide convincing evidence to the human checker. That might diminish the need to use subjective evidence from replies or propagation networks, which is widely leveraged in recent research studies for rumor verification, e.g., [5]–[8].

Traditionally, the propagation networks are leveraged in most of the existing studies for rumor verification in social media such as the structure of replies [6], [7], [9]–[12], stance of replies [4], [8], [12]–[17], or metadata of retweeters [18]. In addition to the propagation networks, incorporating evidence from the Web was proposed by Dougrez-Lewis, Kochkina, Arana-Catania, *et al.* [19] and Hu, Guo, Chen, *et al.* [20]. Figure 1.1 shows the traditional pipeline for rumor verification using the existing proposed evidence sources. Given a rumor expressed in a tweet, both the replies thread and relevant Web articles are retrieved. Evidence sentences are then retrieved from the Web articles to be exploited along with the replies structure and the replies stance in addition to other potential signals by the rumor verification model to decide the veracity of the rumor.

Arabic is one of the most popular languages in Twitter [21], yet under-explored

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

[2] https://datareportal.com/reports/digital-2023-july-global-statshot

[3] https://www.poynter.org/ethics-trust/2018/americans-believe-two-thirds-of-news-on-social-media-is-misinformation/

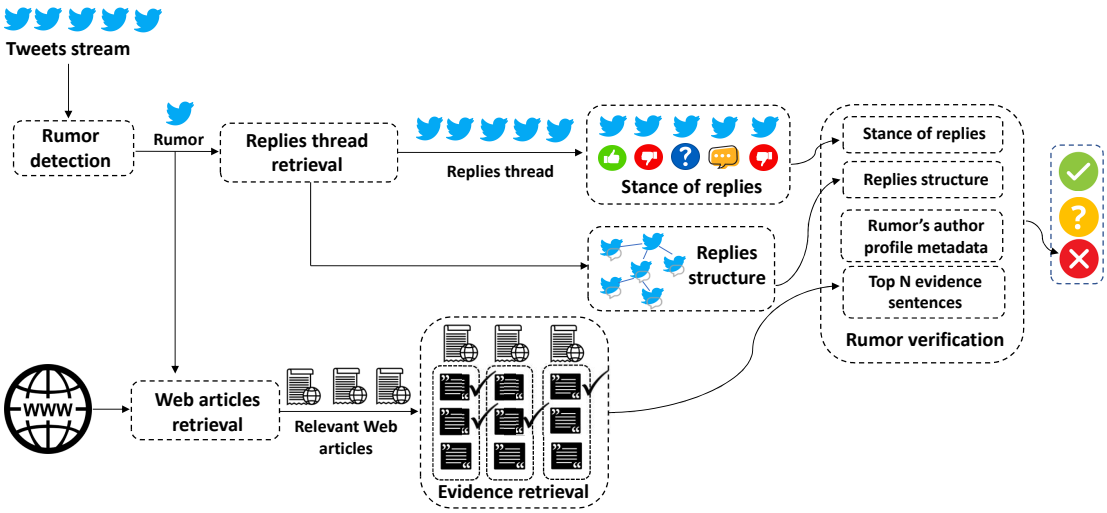



Figure 1.1. A traditional rumor verification pipeline.

for rumor verification. Previous studies have almost exclusively utilized the tweet textual content for verification [22]–[27]. Recently, Haouari, Hasanain, Suwaileh, *et al.* [7] leveraged the conversation thread structure, Althabiti, Alsalka, and Atwell [28] incorporated the detected sarcasm and hate speech in the replies, while Albalawi, Jamal, Khadidos, *et al.* [29] exploited the images and videos embedded in the tweet to address the problem.

To the best of our knowledge, no previous research has incorporated authorities (i.e., entities having the real knowledge or power to verify or deny a specific rumor) for rumor verification. We believe they can 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. To bridge this gap, in this dissertation, we tackle the problem of rumor verification using evidence from authorities. This includes finding authority Twitter accounts that can help verify a given rumor circulating in Twitter, and utilizing both stance of authority tweets and evidence tweets retrieved from their timelines for rumor verification.

We provide an overview of our proposed research problem in Section 1.1. We briefly summarize the proposed solutions and present the major findings in Section 1.2 and 1.3 respectively. We present the contributions of this dissertation in Section 1.4. Finally, we list our research outcomes in Section 1.5.

## 1.1. Research Problem

In this dissertation, we propose and address the problem of *rumor verification using evidence from authorities over Twitter* problem. To show our perception of the role of authorities in the rumor verification pipeline, we present in Figure 1.2 a high-level overview of the full pipeline where we augment the traditional pipeline (presented in Figure 1.1) with our proposed components. Given a rumor expressed in a tweet, the corresponding authority Twitter accounts are retrieved. Evidence tweets are then retrieved from the authorities timeline and the stance of their tweets is detected to be exploited along with other sources of evidence by the rumor verification model to decide the veracity of the rumor. To address our proposed problem we decompose it into a

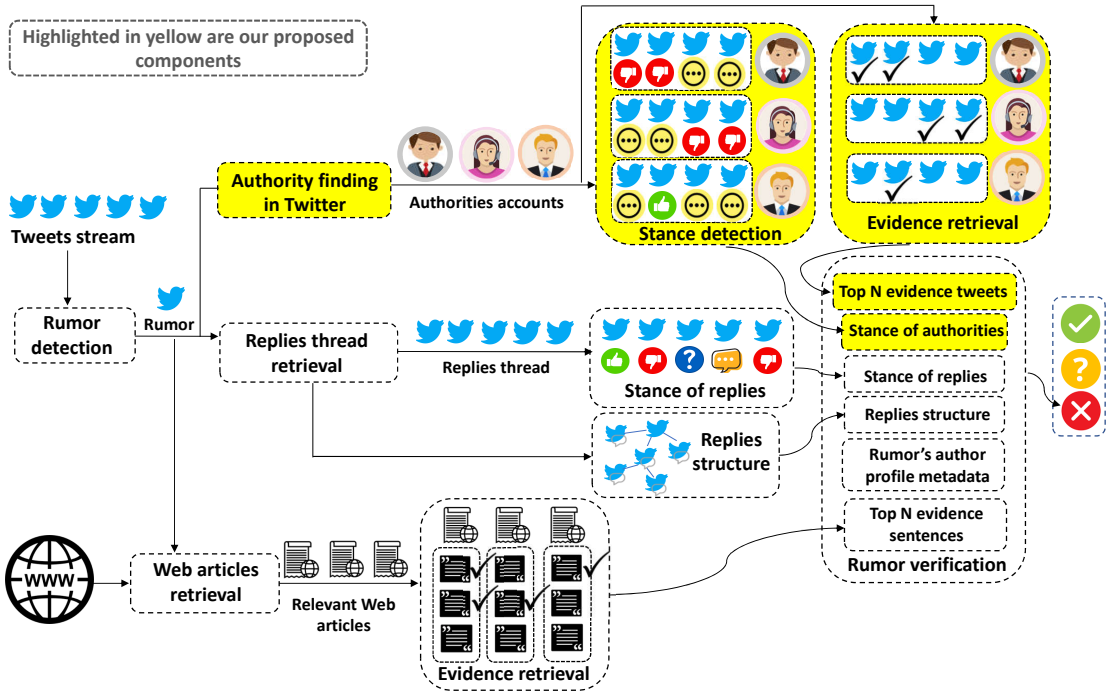



Figure 1.2. Our augmented rumor verification pipeline.

pipeline of four sub-problems defined as follows:

1. **Authority Finding in Twitter:** Given a tweet stating a rumor, retrieve a ranked list of authority accounts from Twitter that can help verify the rumor, i.e., they may tweet evidence that supports or denies the rumor.

2. **Detecting Stance of Authorities towards Rumors:** Given a rumor expressed in a tweet and a tweet posted by an authority of that rumor, detect whether the tweet *supports* (agrees with) the rumor, *denies* (disagrees with) it, or *not* (other).

3. **Evidence Retrieval from Authorities:** 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, retrieve the top N evidence tweets from those timelines.

4. **Rumor Verification using Evidence from Authorities:** Given a rumor expressed in a tweet and a set of retrieved evidence tweets from authorities Twitter accounts for that rumor, the system should determine if the rumor is supported, refuted, or unverifiable.

Accordingly, we aim to address the following high-level research questions:

- **RQ1:** Can our proposed model for authority finding outperform existing topic expert finding models adapted to our task? (refer to Chapter 3)

- **RQ2:** How will stance models trained with existing stance towards Arabic claims datasets perform on the task of detecting the stance of authorities? and will these datasets improve the performance of the models compared to using our dataset only? (refer to Chapter 4)

- **RQ3:** How effective are existing evidence retrieval for fact-checking models on evidence retrieval from authorities? (refer to Chapter 5)
- **RQ4:** How effective are existing claim verification for fact-checking models on rumor verification using evidence from authorities? (refer to Chapter 6)

## 1.2. Overview of Proposed Solution

In this dissertation, we introduce the problem of *rumor verification using evidence from authorities over Twitter* which we decompose into a pipeline of four sub-problems (refer to Section 1.1) namely 1) *authority finding in Twitter*, 2) *detecting the stance of authorities towards rumors in Twitter*, 3) *evidence retrieval from authorities* and 4) *rumor verification using evidence from authorities*.

Our exploration on existing datasets for rumor verification in social media highlighted the need for multiple datasets to address our proposed research problem. Thus, we constructed and released three datasets targeting the *Arabic* language namely 1) the first Authority FINding in Twitter (AuFIN)[4], 2) the first Authority STance towards Rumors (AuSTR)[5], 3) the first Authority-Rumor-Evidence Dataset (AuRED).[6]

To address *authority finding in Twitter* sub-problem (refer to Chapter 3), we propose a hybrid retrieval model that combines lexical and semantic signals in addition to users profiles and network features to find authorities given a rumor. Furthermore, we study the effect of expanding the rumor by exploiting external knowledge bases on the performance of our authority finding model.

To address *detecting the stance of authorities towards rumors in Twitter* sub-problem (refer to Chapter 4), we investigate the usefulness of existing Arabic stance datasets towards claims for our task. Moreover, we explore the mitigation of the traditional class-imbalance issue in stance datasets by experimenting with various loss functions.

To address *evidence retrieval from authorities* sub-problem (refer to Chapter 5), we adopted state-of-the-art (SOTA) evidence retrieval for fact-checking models to investigate their performance on our task. Given that these existing models were originally proposed to retrieve evidence from English Wikipedia pages for English free-text claims, and that both our rumors and authority timelines are Arabic tweets, we adapted the existing models to work on Arabic tweets. Moreover, we study the transfer potential of existing fact-checking datasets on the task, and explore the usefulness of detecting stance of authorities toward rumors for the evidence retrieval.

Finally, we address *rumor verification using evidence from authorities* (refer to Chapter 6) by investigating how existing SOTA claim verification for fact-checking models perform on our task by adapting them to our data. Similar to evidence retrieval, we study the transfer potential of existing fact-checking datasets on the task.

## 1.3. Overview of Research Findings

In this section, we elaborate on the main findings of this dissertation proposing a new potential source of evidence for better rumor verification in social media. Our

[4] https://github.com/Fatima-Haouari/AuFIN
[5] https://github.com/Fatima-Haouari/AuSTR
[6] https://github.com/Fatima-Haouari/AuRED

research findings are as follows:

1. Our exploration on existing datasets for rumor verification highlighted the need for datasets to address our proposed problem. Thus, we constructed and released three datasets namely AuFIN, AuSTR, and AuRED to address the problem and encourage the research community to work on it. We targeted Arabic as it is one of the most used languages in Twitter, yet it is under-studied for rumor verification. However, for all datasets we shared our language independent construction and annotation guidelines to encourage the construction of similar datasets in other languages.

2. Our experimental results suggest that Twitter lists and network features namely the followers, followees, and Twitter lists count, adopted previously for topic expert finding models, play a crucial role in authority finding; however, they are insufficient. This motivates the need to explore other features to differentiate experts from authorities. Moreover, our proposed hybrid model incorporating lexical, semantic, and user network features achieved a modest performance, 0.41 as precision at depth 1, which indicates that finding authorities is a challenging task, and that there is still room for continued enhancement.

3. Our results also highlighted that adopting existing Arabic stance datasets for claim verification is somewhat useful but clearly insufficient for detecting the stance of authorities. Moreover, we found that AuSTR solely, despite the limited size, can be sufficient for detecting the stance of authorities achieving a performance of 0.84 macro-F1 and 0.78 F1 on debunking tweets. Moreover, our experimental results demonstrated that detecting the stance of authorities can be useful for evidence retrieval from authorities.

4. Our investigation on the effectiveness of existing fact-checking (claim verification using evidence from Wikipedia pages) models on our problem highlighted that although evidence retrieval for fact-checking models perform relatively well on evidence retrieval from authorities establishing strong baselines, 0.70 as recall at depth 5, there is still a big room for improvement. However, existing claim verification for fact-checking models perform poorly on rumor verification using evidence from authorities, 0.42 as macro-F1, no matter how good the retrieval performance is. Moreover, existing fact-checking datasets showed a potential in transfer learning to our problem, however, further investigation using different setups and datasets is required.

Throughout the dissertation, we elaborate on these findings and their respective research questions.

## 1.4. Contributions

The contribution of this dissertation target augmenting the traditional rumor verification pipeline by proposing a new source of evidence for better rumor verification. This is achieved by introducing a new problem, constructing and releasing multiple datasets to facilitate future research on the new problem, and learning models addressing the problem.

1. We introduce the new problem of *rumor verification using evidence from authorities over Twitter*. To address the problem we decompose it into four sub-problems namely 1) *authority finding in Twitter*, 2) *detecting the stance of authorities towards rumors in Twitter*, 3) *evidence retrieval from authorities*, and 4) *rumor verification using evidence from authorities.*

2. To address our new proposed problem, we construct and release three datasets namely 1) the first Authority FINding in Arabic Twitter (AuFIN), 2) the first Authority STance towards Rumors (AuSTR) , 3) the first Authority-Rumor-Evidence Dataset (AuRED).

3. We propose a hybrid authority finding model that incorporates both the lexical and semantic relevance in addition to the users' network features.

    - We explore the effect of rumor expansion by exploiting Knowledge Bases on the performance of our proposed model.
    - We conduct a thorough error analysis on our proposed hybrid model to gain insights for future improvements.
    - We deploy our proposed authority finder model into a real-time Arabic claim verification system.

4. We investigate the usefulness of existing Arabic datasets of stance towards claims for detecting the stance of authorities towards rumors.

    - We explore the adequacy of existing Arabic datasets of stance towards claims for the task, and the effect of augmenting our in-domain AuSTR data with those datasets on the performance of the stance models.
    - We investigate the performance of the stance models when adopting variant loss functions to alleviate the class-imbalance issue.
    - We perform a thorough failure analysis to gain insights for future work on the task.

5. We investigate the usefulness of detecting stance of authorities toward rumors for evidence retrieval from authorities.

6. We investigate the effectiveness of existing fact-checking models on rumor verification using evidence from authorities.

    - We explore the effectiveness of evidence retrieval for fact-checking models on evidence retrieval from authorities.
    - We study the effectiveness of claim verification for fact-checking models on rumor verification using evidence from authorities.
    - We explore the transfer potential of existing fact-checking datasets to evidence retrieval and rumor verification using the retrieved evidence from authorities tasks.

7. We co-organize our proposed sub-problems as shared tasks to motivate the research community to work on our proposed problem.

- We co-organize the *authority finding* sub-problem as a task in CheckThat! Lab at CLEF 2023.
- We co-organize the *evidence retrieval from authorities*, and *rumor verification using evidence from authorities* sub-problems as tasks in CheckThat! Lab at CLEF 2024.

## 1.5. Research Outcomes

This dissertation resulted in multiple publications, a system, and several shared tasks.

### *1.5.1. Publications*

- Journal Articles:
  - **Fatima Haouari**, Tamer Elsayed. Are Authorities Denying or Supporting? Detecting Stance of Authorities Towards Rumors in Twitter. Social Network Analysis and Mining. 2024. [30]
  - **Fatima Haouari**, Tamer Elsayed, Watheq Mansour. Who Can Verify This? Finding Authorities for Rumor Verification in Twitter. Information Processing and Management. 2023. [31]
- Conference Papers:
  1. **Fatima Haouari**, Tamer Elsayed. Detecting Stance of Authorities towards Rumors in Arabic Tweets: A Preliminary Study. Proceedings of the 45rd European Conference on Information Retrieval (ECIR 2023). [32]
  2. Zien Sheikh Ali, Watheq Mansour, **Fatima Haouari**, Maram Hasanain, Tamer Elsayed, and Abdulaziz Al-Ali. Tahaqqaq: A Real-Time System for Assisting Twitter Users in Arabic Claim Verification (SIGIR 2023). [33]
  3. **Fatima Haouari**. Evidence-based Early Rumor Verification in Social Media. Proceedings of the 44rd European Conference on Information Retrieval (ECIR 2022). [34]
  4. **Fatima Haouari**, Marwa Essam and Tamer Elsayed. bigIR at TREC 2020: Simple but Deep Retrieval of Passages and Documents. Proceedings of the 29th Text REtrieval Conference (TREC 2020). [35]
- Workshop Papers:
  1. **Fatima Haouari**, Maram Hasanain, Reem Suwaileh and Tamer Elsayed. ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, (WANLP 2021). [7]
  2. **Fatima Haouari**, Maram Hasanain, Reem Suwaileh and Tamer Elsayed. ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, (WANLP 2021). [36]
- Book Chapters:

1. A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, T. Elsayed, D. Azizov, T. Caselli, G. Cheema, **Fatima Haouari**, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouani, Overview of the CLEF–2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. (CLEF 2023). [37]
2. Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, **Fatima Haouari**, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, Yavuz Selim Kartal. Overview of the CLEF–2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. Experimental IR Meets Multilinguality, Multimodality, and Interaction. (CLEF 2021). [38]
3. Alberto Barron-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, **Fatima Haouari**, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. Proceedings of the Eleventh International Conference of the CLEF Association. (CLEF 2020). [39]

- Organization of shared tasks papers:

1. Alberto Barrón-Cedeño, Firoj Alam, T. Chakraborty, Tamer Elsayed, P. Nakov, P. Przyby la, J. Struß, **Fatima Haouari**, Maram Hasanain, Federico Ruggeri, Xingyi Song, Reem Suwaileh. The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. European Conference on Information Retrieval. (ECIR 2024). [40]
2. Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, **Fatima Haouari**, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, Gullal S. Cheema, Dilshod Azizov and Preslav Nakov. The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority. Proceedings of the 45rd European Conference on Information Retrieval (ECIR 2023). [41]
3. Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, **Fatima Haouari**, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. Proceedings of the 43rd European Conference on Information Retrieval (ECIR 2021). [42]
4. Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and **Fatima Haouari**. CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media. Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020). [43]

### *1.5.2. Systems*

Our *Authority Finding* model was deployed as part Tahaqqaq [33], a real-time system for assisting Twitter users in Arabic claim verification, to enable users to find authorities for a given tweet in real-time or any free-text claim.

### *1.5.3. Organization of Shared Tasks*

We co-organized three out of the four sub-problems in our proposed *rumor verification using evidence from authorities* pipeline as shared tasks to motivate the research community to work on our problem. Specifically we co-organized the below:

1. CheckThat! Lab at CLEF 2023: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Sources [37], [41], [44]. In this lab we introduced *Authority Finding in Twitter* sub-problem.

2. The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities and Adversarial Robustness. In this lab [40], [45] we introduced both *evidence retrieval from authorities*, and *Rumor Verification using Evidence from Authorities* sub-problems.

We also co-organized other relevant shared-tasks but not directly related to the proposed research problem as presented below:

1. CheckThat! Lab at CLEF 2020: Enabling Automatic Identification and Verification of Claims in Social Media [22], [39], [43], [46]. In this lab we co-organized the *Check-worthy claims detection*, *evidence retrieval from Web articles for claim verification*, and *claim verification using evidence from Web articles* tasks in Arabic.

2. CheckThat! Lab at CLEF 2021: Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News [38], [42], [47], [48]. In this lab we co-organized both *Check-worthy claims detection*, and *detecting previously fact-checked claims* tasks in Arabic.

The remainder of this dissertation is organized as follows. Chapter 2 presents our extensive literature review. Chapter 3 presents our work on the authority finding sub-problem. Chapter 4 presents our work on detecting the stance of authorities towards rumors. Chapter 5 presents our study for evidence retrieval from authorities. Chapter 6 presents our efforts on rumor verification using evidence from authorities. Finally, in Chapter 7 we conclude the dissertation and presents the implications, limitations, and future directions of our work.

# CHAPTER 2: RELATED WORK

In this section, we present an extensive literature review. We first review studies targeting rumor verification in social media in Section 2.1. We cover evidence-based fact-checking studies in Section 2.2. Finally, we review experts finding in social media studies in Section 2.3.

## 2.1. Rumor Verification in Social Media

In this section, we present the rumor verification in social media studies and discuss their proposed approaches (Section 2.1.1) and datasets (Section 2.1.2).

### *2.1.1. Proposed Approaches*

A plethora of studies addressed rumor verification in social media exploiting variant sources of evidence. In this section, we present the main evidence sources adopted by existing works, then we present the existing proposed solutions.

The widely used source of evidence adopted by existing studies is the propagation networks, i.e., replies and retweets. Additionally user features and evidence from Web articles were also utilized for rumor verification along with other features. We present all the sources utilized by existing studies below:

***Propagation networks structure:*** Exploiting the propagation networks structure was first introduced by Ma, Gao, and Wong [49], motivated by the hypothesis that different types of rumors may have distinct propagation structures. To motivate research on this path, they extended two existing Twitter datasets [50], [51], and released SOTA datasets namely Twitter15 and Twitter16 that contain the propagation tree for each post where posts are labeled as non-rumor, true rumor, false rumor, or unverified.

***Replies stance:*** The stance of replies for rumor verification was first proposed by Zubiaga, Liakata, Procter, *et al.* [4] who analyzed how users react to rumors spreading in social media. It was then adopted by many studies for rumor verification [52]–[56]. To motivate research on the task, rumor verification using stance of replies was also introduced as a shared task [56]–[58].

***User features:*** The user profiles features of retweeters or repliers were adopted by many rumor verification studies [18], [49], [59] studies. Liu and Wu [18] showed that using retweeters profiles only can be used to verify the content of a tweet. Some of the user features adopted are followers and friends count, whether the user account is verified or not, and length of username and her profile description.

***Web articles:*** Recently, Dougrez-Lewis, Kochkina, Arana-Catania, *et al.* [19] proposed augmenting the propagation networks with evidence from the Web. Given a rumor and a relevant sentence from relevant Web pages, they predict the veracity of the rumor. They determine the final label for each rumor using majority voting over top relevant sentences. In the other hand, Hu, Guo, Chen, *et al.* [20] proposed exploiting both text and images retrieved from the Web as sources of evidence.

***Other features:*** Some authors have also suggested other features for better rumor verification. Geng, Lin, Fu, *et al.* [60] proposed including the sentiment of replies. Althabiti, Alsalka, and Atwell [28] proposed detecting sarcasm and hate speech in the replies for Arabic rumor verification in Twitter.

#### *2.1.1.1. Non-neural Models*

Ma, Gao, and Wong [49] trained a kernel-based SVM classifier model that uses propagation tree kernels constructed using posts content, user profiles, and the propagation structure. On the other hand, a graph-kernel based hybrid SVM classifier that captures the propagation patterns and the semantic features such as topics and sentiments was proposed by Wu, Yang, and Zhu [61]. Furthermore, Dang, Moh'd, Islam, *et al.* [62] proposed a Naive Bayes model that utilizes multiple features such as topic, sentiment, network structural features.

#### *2.1.1.2. Neural Networks Models*

Due to the time complexity of computing the similarities between propagation trees in their kernel-based model [49], Ma, Gao, and Wong [9] further proposed using recursive neural networks that catch both content and the propagation structure. Then they enhanced their model with a specific attention mechanism that can capture the most evidential replies to detect the rumor [63]. Yuan, Ma, Zhou, *et al.* [64] proposed a CNN with multi-head attention that integrates global structural and local semantic information based on the heterogeneous propagation network. Furthermore, a multivariate time series using CNN and GRU model was proposed by Liu and Wu [18] who showed that using retweeters profiles only can be used to verify the content of a tweet, but their model does not capture the variational lengths of different propagation trees. On the other hand, Ma, Gao, and Wong [52] proposed a multi-task GRU learning model that detect the stance of comments and the tweet veracity jointly by exploiting text content only. Li, Zhang, and Si [53] extended their work by using user profiles and adopting LSTM with the attention mechanism in their study. Alternatively, Kumar and Carley [15] employed variant models of LSTM namely, branch LSTM and tree LSTM with different embedding inputs using BERT, Glove, and DeepMoji to detect stance of comments and tweet veracity. Differently Liu and Wu [5] proposed using Positive and unlabeled learning (PU-learning) to simulate a real-world scenario where data is unlabeled and imbalanced.

#### *2.1.1.3. Graph-based and Graph Neural Models*

A series of recent studies has proposed using graph neural models to capture the propagation structure of the rumors. Bian, Xiao, Xu, *et al.* [10] showed that graph convolutional networks can capture the propagation structure better than recursive neural networks models proposed by Ma, Gao, and Wong [9]. Huang, Yu, Wu, *et al.* [59] adopted the Graph attention network to construct a tweet-word-user heterogeneous graph based on the text contents and the source tweet propagations of rumors. Lu and Li [65] presented a Graph-aware Co-Attention networks which exploits the co-attention mechanism to provide explainability on source tweets, the retweets propagation, and the retweeters charecteristics. The authors found that to detect false rumors the charceteristics of users who early retweet the rumorous tweet should be examined. On the other hand, Rosenfeld, Szanto, and Parkes [66] proposed a Weisfeiler-Lehman graph kernel to extract complex topological information from the propagation networks. Differently, Graph Adversarial Contrastive Learning was proposed by Sun, Qian, Dong, *et al.* [67] to fight complex cases such as noise and adversarial rumors. They introduced contrastive learning as part of the loss function for explicitly differentiating between replies threads

of the same class and different classes. Some authors have also suggested considering the temporal evolution of the propagation networks [6], [11], [68], and addressed the problem utilizing temporal-aware graph neural networks.

#### *2.1.1.4. Transformers and pre-trained Language Models*

Contextualized transformers-based models such as BERT [69] have shown superiority in different Natural Language Processing tasks and consequently they were adopted for rumor verification. Ma and Gao [70] proposed a tree transformer model for early verification of rumors. The model exploits the replies content and their structure. The authors further enhanced it with the attention mechanism in order to visualize the most important words in the replies to verify the rumor in the source tweet. Tian, Zhang, Wang, *et al.* [71] addressed early rumour verification based on only the tweet content and its early user replies. They proposed CNN- and BERT-based deep neural models to learn attitude representation from user replies via transfer learning. They further proposed CNN-BiLSTM and BERT neural models to integrate attitude representation and content representation for tweets and their replies. Additionally, the multi-head attention mechanism in a transformer network was adopted by Khoo, Chieu, Qian, *et al.* [72] to model the long distance interactions between posts where a user reply to the entire conversation thread rather than a specific user. Alternatively, Yu, Jiang, Khoo, *et al.* [17] proposed a hierarchical BERT model for stance classification and rumor verification as a single task. Another study that adopted BERT for stance detection for rumor verification is the one done by Radhakrishnan, Kanakagiri, Chakravarthy, *et al.* [55]. Longformer [54], a transformer-based model for long documents, was exploited by Khandelwal [56]. The authors used the source tweet and its replies to detect the stance of replies towards the source and the veracity of the source jointly. The source and all replies were given as an input to the longformer where they added a [CLS] and [SEP] tokens before and after each post respectively.

### *2.1.2. Datasets*

As presented in Table. 2.1, we review existing rumor verification in social media datasets. As shown in the table, most of existing studies focus on using the propagation networks as evidence. Some of the studies relied on the rumor textual content solely without any external evidence for verification [23], [29], [73], [74]. Recently, some studies incorporated evidence from the Web such as relevant Web articles [19], [22], [75] and images [75] in addition to social media users' metadata [76]. Most of the existing datasets for Arabic Rumor verification do not incorporate any external evidence. Some notable exceptions are the data released by Haouari, Hasanain, Suwaileh, *et al.* [7] and Hasanain, Haouari, Suwaileh, *et al.* [22] who incorporated the propagation networks and Web articles as external evidence respectively. Compared to existing datasets, our constructed rumor verification dataset incorporates evidence from authorities Twitter timelines (refer to Chapter 6).

## 2.2. Evidence-based Fact-Checking

In this section, we present the fact-checking studies and discuss their proposed approaches (Section 2.2.1) and datasets (Section 2.2.2).

Table 2.1. Existing datasets for rumor verification in social media.

| Dataset | # Rumors | Platform/Lang | Evidence |
|---|---|---|---|
| Arabic-COVID19 [73] | 2,000 | Twitter/Ar | None |
| Multimodal-Rumors [29] | 4,025 | Twitter/Ar | None |
| COVID-19-FAKES [23] | 220K | Twitter/Ar+Eng | None |
| PHEME [4] | 330 | Twitter/En | Propagation networks |
| RumorEval17 [58] | 325 | Twitter/En | Propagation networks |
| RumorEval19 [57] | 446 | Twitter+Reddit/En | Propagation networks |
| Twitter15/16 [49] | 818 | Twitter/Eng | Propagation networks |
| Weibo [51] | 4,664 | Weibo/Zh | Propagation networks |
| DAST [77] | 220 | Reddit/Da | Propagation networks |
| ArCOV19-Rumors [7] | 3,584 | Twitter/Ar | Propagation networks |
| CheckThat!2020 [22] | 165 | Twitter/Ar | Web articles |
| PHEMEPlus [19] | 1972 | Twitter/En | Propagation networks/Web articles |
| MuMIN [76] | 12,914 | Twitter/Multi | Propagation networks/Metadata |
| $MR^2$ [75] | 14,700 | Twitter +Weibo/En+Zh | Propagation networks/Web articles and images |

### *2.2.1. Proposed Approaches*

Fact-checking using evidence from Wikipedia pages was introduced as part of FEVER shared task by Thorne, Vlachos, Cocarascu, *et al.* [78]. The task is a pipeline of three subtasks namely relevant documents retrieval, evidence retrieval, and claim verification using the retrieved evidence. A plethora of studies addressed the task contributing to one of the subtasks or multiple subtasks.

#### *2.2.1.1. Document Retrieval Models*

Most of the fact-checking studies [79]–[84] adopted the entity linking approach proposed by Hanselowski, Zhang, Li, *et al.* [85] for document retrieval. The approach is to extract all phrases which potentially indicate entities from the claim, and utilize them as search queries for retrieval. On the other hand, Jiang, Pradeep, and Lin [86] proposed adopting both BM25 and the entity linking approach for retrieval. They combine the output of both approaches by alternating through the two ranked lists of documents, then they deduplicate the documents to keep the top K unique ones. Recently, Chen, Zhang, Guo, *et al.* [87] proposed a generative document retrieval model where they adopted a pre-trained sequence-to-sequence model to generate documents titles. Inspired by the entity linking approach [85], DeHaven and Scott [88] extracted entities from the claim however they used a fuzzy string search system to retrieve relevant documents.

*2.2.1.2. Evidence Retrieval Models*

Existing studies either adopted a keyword matching approach [89], neural ranking models [79], [85], [90], [91], or pre-trained models [82], [83], [86]–[88], [92]–[94] for evidence retrieval. Luken, Jiang, and Marneffe [89] used all nouns and named entities in the claim to match against all sentences. Hanselowski, Zhang, Li, *et al.* [85] enhanced an existing Bi-LSTM based model proposed originally for natural language inference (NLI) task. They specifically extended the model to generate a ranking score given two input sentences, instead of predicting their entailment relation. This model was then further enhanced by Zhou, Han, Yang, *et al.* [79] who added a relevance filter to exclude sentences with relevance scores below the threshold. Bi-LSTM was also adopted by Nie, Wang, and Bansal [91] in their proposed Neural Semantic Matching Network.

Recent studies, focused primarily on adopting pre-trained models such as BERT [95], but with variant loss functions and some additional enhancements. Some addressed the task as a binary classification task [82], [86], [88], [94], others proposed a pairwise ranking model [92], [93], while others explored distance-based loss functions [96]. Differently, Chen, Zhang, Guo, *et al.* [87] proposed a generative evidence retrieval model to ensure retrieving precise set of evidence sentences, where they adopted BART [97] to generate relevant documents and evidence identifiers.

*2.2.1.3. Claim Verification Models*

Most of the existing works target the claim verification step and adopt existing approaches for evidence retrieval. Many studies formulated the claim verification task as an NLI task. i.e., requires a model to predict the relationship between a pair of premise and hypothesis as entailment, contradiction or neutral. They approached the task with different NLI methods [89], [98], [99]. A claim and the concatenated set of evidence was simply fed to an NLI model by Thorne, Vlachos, Cocarascu, *et al.* [78]. Hanselowski, Zhang, Li, *et al.* [100] and Nie, Chen, and Bansal [101] adopted the enhanced LSTM inference model (ESIM) [102] where they used the max pooling to aggregate the information from the concatenated evidence sentences.

Pre-trained transformers were also exploited for the task [92], [103]. Given the top 5 evidence sentences already selected, Soleimani, Monz, and Worring [92] fine-tuned BERT [95] to predict the veracity of the claim given an evidence sentence. The final label was decided by aggregating the five scores. A similar approach was adopted by Stammbach and Neumann [103] but they concatenated the Wikipedia page title to each evidence sentence for co-reference resolution. Jiang, Pradeep, and Lin [86] fine-tuned T5 [104] for claim verification. They proposed a listwise approach, where all evidence sentences of a claim are considered together. Moreover, they proposed a data augmentation technique to introduce noise in the training data. Recently, DeHaven and Scott [88] proposed a mixed approach that combines the prediction of two models. The first model considers only claim-evidence pairs while the second considers claim and all evidence sentences as an input to the model. The scores of these models are then aggregated to get the final verification label.

Recently, graphs [79] and graph neural networks [105]–[107] were introduced to capture the semantic relations between evidence sentences. Zhou, Han, Yang, *et al.* [79] proposed a graph-based evidence reasoning system for claim verification where they modeled the set of evidence sentences as a fully connected graph to propagate information among the evidence set. Similarly, Liu, Xiong, Sun, *et al.* [80] adopted

the fully connected graph to represent the evidence sentences but incorporated two sets of kernels to learn a fine-grained evidence representation in addition to adopting the graph attention for aggregation to jointly reason the graph. With the aim to employ the semantic representation at both word and graph level during the reasoning process, Zhong, Xu, Tang, *et al.* [108] applied semantic role labeling to parse each evidence sentence, and establish links between arguments to construct the graph. Then they adopted graph convolutional network and graph attention network to propagate and aggregate information over the graph structure.

#### *2.2.1.4. Joint Evidence Retrieval and Claim Verification Models*

A few studies proposed joint models for evidence selection and claim verification. Pointer networks to jointly extract evidence and verify the claim were adopted by Hidey and Diab [109] and Hidey, Chakrabarty, Alhindi, *et al.* [110]. Yin and Roth [111] proposed a joint framework consisting of two convolutional neural network for each task and a shared representation of the claim. Differently, Si, Zhou, Li, *et al.* [82] adopted the capsule networks to capture the implicit stance of the evidence towards the claim. Additionally, a reinforcement learning approach was proposed by Wan, Chen, Du, *et al.* [81] who used the deep Q-learning to identify the candidate evidences, then refine those candidate with a post-processing strategy. Moreover, a multi-attention model that enables both sentence and token self-attention was proposed by Kruengkrai, Yamagishi, and Wang [112]. Different than other studies who either concatenate all evidences, or process each claim-evidence pair separately, Subramanian and Lee [113] proposed a framework for extracting evidences sets. The claim verification is then performed hierarchically based on each evidence set, and then based on all evidence sets.

### *2.2.2. Datasets*

A plethora of datasets for claim verification using evidence from Wikipedia were constructed and released to motivate research on the task. The first introduced dataset is FEVER [78], which consists of 185,445 English claims generated by modifying sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were generated from. The claims are labeled as Refuted , Supported, or NOT Enough Info to verify them. HoVer dataset was then released by Jiang, Bordia, Zhong, *et al.* [114] for many-hop evidence extraction and fact verification to challenge systems to retrieve evidence from multiple Wikipedia articles. To address the limitation of the synthetic nature of claims in FEVER, Sathe, Ather, Le, *et al.* [115] constructed and released WikiFactCheck-English which consists of a large collection of real claims.

Another group of datasets was introduced to allow evidence to be extracted from tables in Wikipedia pages, e.g., TabFact [116] and INFOTABS [117]. Furthermore, to include evidence from both text and tables, FEVER dataset was extended by Aly, Guo, Schlichtkrull, *et al.* [118] introducing FEVEROUS dataset.

Other researchers presented non-English datasets for the task such as Dan-FEVER [119] which consists of 6,407 Danish claims, and CsFEVER [120] which is a Czech version of FEVER.

## 2.3. Expert Finding in Social Media

The expert finding problem was addressed in many domains [121]–[123] including community question answering platforms [124]–[130], academic social networking sites [131], and micro-blogging platforms [132]–[145]. In this section we review expert finding in social media studies which can be categorized into three main categories:

### *2.3.1. Topic Experts Finding*

Given a topic, the task is to find a ranked list of experts [123]. Some traditional approaches [132], [138] exploited the followers relations, user profiles and tweets to determine the topic of influence of users. Weng, Lim, Jiang, *et al.* [132] proposed TwitterRank for finding topic influencers by exploiting their Twitter follower graph. Pal and Counts [138] adopted the Gaussian Mixture Model [146] to find potential experts for a given topic. Ghosh, Sharma, Benevenuto, *et al.* [139], on the other hand, introduced exploiting Twitter lists created by users to identify topical experts. Additionally, Bozzon, Brambilla, Ceri, *et al.* [140] ranked users by their cumulative informative from multiple social networks. Yeniterzi and Callan [141] proposed topic-dependent sub-graphs, which are constructed with the users activities on topic-relevant posts. Wei, Cong, Miao, *et al.* [142] exploited both followers relations and Twitter lists for the task. Moreover, Lahoti, De Francisci Morales, and Gionis [143] utilized the users Twitter lists solely and proposed a variant query-dependent personalized PageRank for the task. Differently, Horne, Nevo, and Adalı [144] formulated the problem as a classification task where they proposed a model to predict whether a user is an expert or not for a given topic in Twitter and Reddit social media platforms.

### *2.3.2. Local Experts Finding*

Local experts, compared to topic experts, are experts in a topic around a specific location. Existing studies [133]–[135], [145], [147] exploited geo-tagged Twitter lists to address the problem. Cheng, Caverlee, Barthwal, *et al.* [145] proposed decomposing the local expertise of a given user into a topical expert and a local expert and then combining them in a linear way to arrive to the final score. Niu, Liu, and Caverlee [133], [134], on the other hand, proposed addressing the problem using a supervised learning-to-rank framework to rank candidate local experts. Additionally, Li, Eickhoff, and Vries [135] proposed multiple probabilistic models to find local experts in Twitter. Local experts finding across multiple platforms was addressed by Ma, Yuan, Wang, *et al.* [147] who proposed dividing the problem into sub-problems, and exploiting the common users as bridges to transfer knowledge.

### *2.3.3. Misinformation-based Experts Finding*

To our knowledge, only two studies targeted finding the most relevant experts for a piece of misinformation or rumor [136], [137]. Liang, Liu, and Sun [136] targeted the Chinese Weibo social media platform. The authors proposed a tag-based approach that relies on the tags users assign to themselves in Weibo and usually represents their expertise. They estimated the probability that a user is an expert using a Bayes theorem given the tags in the user profiles and the tags predicted for the post by their trained tag suggestion model. Similar to [139], Li, Dong, Yang, *et al.* [137] took advantage

of Twitter lists in their work, and they represented each user as a document containing his profile information, posted tweets, and metadata of lists he is a member of. They employed a list topic model to map users' information, then they used a composite model to get a matching score between the misinformation and the user, and similar to [136], they adopted the Bayes theorem to estimate the probability that a user is an expert. Neither of the two studies constructed nor released a test collection for the task, and both relied on hiring annotators to get relevance judgments given the output of their systems.

# CHAPTER 3: AUTHORITY FINDING FOR RUMOR VERIFICATION

Fact-checkers, or even normal users, who attempt to verify a rumor over social media try to find a trusted source of evidence (relevant to that rumor) that can help them confirm or deny that specific rumor. Authorities (i.e., entities having the real knowledge or power to verify or deny a specific rumor) can 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. Thus, having an automated system for finding authority accounts from Twitter for a given rumor would be a great asset in that regard. In this chapter, we target the first sub-problem in our proposed pipeline namely the *authority finding in Twitter*.

This chapter starts with defining the *authority finding in Twitter* task in Section 3.1. Next, we present a comparison between the authority finding and the existing expert finding task in Section 3.2. In Section 3.3, we present an overview about this work. We describe our test collection construction approach in Section 3.4. Our proposed authority finder model is detailed in Section 3.5. We discuss the experimental setup in Section 3.6 and thoroughly analyze the results, answer the research questions, and discuss the lessons we learned in Section 3.7. We conduct error analysis to gain insights for future work, and discuss the limitations of our study in Section 3.8. We present our proposed model deployment in a real-time Arabic claim verification system in Section 3.9. Finally, we conclude in Section. 3.10.

## 3.1. Problem Definition

We propose the task of *authority finder in Twitter* defined as follows: Given a tweet stating a rumor, retrieve a ranked list of authority accounts from Twitter that can help verify the rumor, i.e., they may tweet evidence that supports or denies the rumor.

## 3.2. Authority Finding vs. Expert Finding

We consider the authority finding problem as a sub-problem of *topical expert finding* defined as "efficiently identifying the right individual (or group) from a field of candidates that has the expertise to provide desired information or complete a desired task" [148]. Therefore, we exhaustively review the efforts made on topic expert finding task in Section 2.3.1 and expert finding in social media in general in Section 2.3. In fact, every authority is considered an expert, but not every expert is an authority.

As presented in Figure 3.1, for the rumor stating that the Sultanate of Oman recalled its ambassador in Qatar, experts or journalists in the Arabian Gulf countries in general or in Qatar or Oman in particular are not considered authorities. However, the embassy of Oman in Qatar, the Qatari and Omani foreign ministers/ministries, the spokesman for ministries, and government communication offices are all considered authorities, as we believe they are the trusted sources to help verify the rumor.[1] Previous research showed that even highly reputable Twitter users, such as news agencies, tend to spread rumors [4], or are at least biased. In our work, we do not consider news agencies as authorities unless the given rumor is about them, or the entities mentioned in the rumor are related to them, e.g., for the rumor "The sports journalist Hafid Derradji was fired from beIN SPORTS", beIN SPORTS channel is indeed an authority.

The expert finding problem was addressed in many domains [121]–[123] including community question answering platforms [127]–[130], academic social networking

[1] That rumor has 9 authorities as per our annotators.


Figure 3.1. Authorities versus experts.

sites [131], and micro-blogging platforms [136], [137], [139], [143]. A closer look at the literature on expert finding in social media reveals that there is no publicly available dataset for expert finding in social media that associates an expert to a given query. Prior studies relied on evaluating their systems by conducting a user study [138], hiring annotators to get relevance judgments given the output of their systems [136], [139], [143], adopted a pooling strategy to get top experts by SOTA expert finding algorithms [137], [141], or constructed a test collection to train and evaluate their models but it was not made publicly available [144].

Expert finding for a relevant piece of *misinformation* is understudied; the literature shows that only two studies addressed this problem [136], [137]. Liang, Liu, and Sun [136] identified experts for Chinese rumors by searching a collection of Sina Weibo users. Their approach relies on "expertise tags" that users assign to themselves. However, we argue that this might not be reliable, as the provided expertise information can be misleading in some cases [149]. Li, Dong, Yang, *et al.* [137] targeted only domain-specific misinformation, which rarely mention specific people, places, or organizations, as per Liang, Liu, and Sun [136] empirical analysis, and evaluated their approach using a small set of only 20 queries. Neither of the two studies constructed nor released a test collection for expert finding for rumor/misinformation verification. Our literature review also reveals that, although several studies on topic and misinformation expert finding in social media have been conducted, there is no single study that explores the use of pre-trained language models or investigates query expansion by exploiting Knowledge Bases (KBs) for expert finding in Twitter.

## 3.3. Overview of Our Work

To the best of our knowledge, there is no work that addresses the expert finding problem in Arabic Twitter in general, and authority finding for rumor verification in particular. In fact, authority finding specifically for rumor verification was never addressed in any language. To fill this literature gap, we first define the problem of authority

finding in Twitter. We then construct the first test collection for the task, release it for future research, and share the guidelines for constructing it to help apply to other languages. Moreover, we propose a hybrid retrieval model that combines lexical and semantic signals in addition to user profile and network features to find authorities given a rumor. Furthermore, we study the effect of expanding the rumor by exploiting external knowledge bases on the performance of our authority finding model. Finally, we deploy our proposed model as part of a real-time system for assisting Twitter users in Arabic claim verification. The contributions of this work are as follows:

1. We introduce and define the problem of authority finding in Twitter.
2. We present the first study addressing authority finding for Arabic rumors.
3. We construct and release[4] the first test collection for the Authority FINding in Arabic Twitter (AuFIN) and share our language-independent construction and annotation guidelines.
4. We propose a hybrid authority finding model that incorporates both the lexical and semantic relevance in addition to the users' network features. We also explore rumor expansion by exploiting Knowledge Bases.
5. We conduct a thorough error analysis on our proposed hybrid model to gain insights for future improvements.
6. We deploy our proposed authority finder model into a real-time Arabic claim verification system.

## 3.4. Constructing a New Test Collection: AuFIN

To address the lack of authority finding test collections, in this work, we introduce the first Authority FINding in Arabic Twitter (AuFIN) test collection for rumors spreading over Arabic Twitter. We target Arabic as it is one of the most dominant languages in Twitter [150], yet it is under-studied for rumor verification. The test collection consists of (1) a set of 150 rumors expressed in tweets, constituting the *query set*, (2) a collection of 395,231 Twitter users, along with Twitter lists they are members of, and their recent timeline tweets,[2] constituting the *document set*, and (3) a manually-annotated list of authority Twitter accounts for each of the rumors, constituting the *relevance judgment set*. In this section, we present our approach of constructing the test collection.

### *3.4.1. Rumors (Queries)*

We selected 150 rumors from Misbar, an Arabic fact-checking website[3] used by previous studies to construct datasets for Arabic rumor verification [7] and articles credibility [151], [152]. Several studies addressed the number of queries in a test collection for reliable evaluation of retrieval systems; Jones and Van Rijsbergen [153] recommended 75 queries, while Buckley and Voorhees [154] found that 50 queries can lead to stable evaluation. Moreover, several TREC test collections, e.g., [155], and existing Arabic test collections, e.g., [156], [157], provided 50 queries for evaluation.

[2]The timeline was collected by the end of November 2021.

[3]`https://misbar.com/`

Accordingly, to select our rumors, we focused on three different categories of rumors, namely, *sports*, *health*, and *politics*, and selected 50 rumors from each to maintain balance. For each category, we focused only on rumors that have, in the corresponding fact-checking article, associated tweets spreading them, and selected one tweet example among them.

### *3.4.2. User Collection*

To construct our user collection, we first collected a seed of users by adopting either Twitter streaming or searching; then, we further expanded the retrieved seed by exploiting their followees and members of their Twitter lists. Finally, for each user in our collection, we collect their social data (timeline and lists), which can be used to extract signals about their expertise and authoritativeness. We present below our approach to gathering the users collection in detail.

#### *3.4.2.1. Seed Users*

To collect a seed of potential authorities Twitter accounts, we adopted two different techniques:

- **Streaming users:** We used the Twitter academic filtered stream API[4] to stream Arabic tweets that match specific keywords/phrases in the user profile name or description that might indicate authoritativeness, e.g., "ministry", "embassy", "organisation", "vice president", "official account", "politics", "sports", "health", etc. Specifically, we used 448 keywords/phrases devised by graduate students[5]. We ran the streaming process continuously for seven days. It is worth noting that although we are targeting Twitter accounts that tweet in Arabic, we included both Arabic and English keywords, because profile names and descriptions are not necessarily written in Arabic, e.g., the U.S. Embassy in Qatar,[6] and the Egyptian Ministry of International Cooperation[7] Twitter accounts. This approach yielded 156,203 unique users, where 3,864 are verified Twitter accounts.

- **Searching users:** Given that we may have missed some key users by relying on streaming only, as they may not have tweeted during our streaming time, we used the Twitter user search API,[8] which retrieves users from Twitter user database. We used the same keywords and phrases but in Arabic only. We were able to use English keywords for streaming because we constrained it to Arabic tweets. However, for searching users, we do not have the option to limit it to Twitter accounts that tweet in Arabic. It is worth noting that we eventually filtered the user accounts that do not have Arabic terms in neither their bio, bio-name, nor last tweet.

After deduplication, our final seed set of users has 182,734 unique users with 5,222 verified accounts.

[4] https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/integrate/build-a-rule#list

[5] we release Keywords used for searching/streaming with our data.

[6] https://twitter.com/USEmbassyDoha

[7] https://twitter.com/moicegypt

[8] https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/api-reference/get-users-search

*3.4.2.2. Users Expansion*

To further expand our seed users collection, we performed the following:

- **Twitter followees:** Assuming that authorities usually follow other authorities, we collected the followees of our *verified* users and further collected the followees of their followees, using the follows lookup Twitter API.[9]

- **Twitter lists:** Given that users that share the same knowledge or expertise are usually added to the same Twitter lists [139], we retrieved the Twitter lists that our *verified* seed users are members of, then we retrieved all the members of those lists, using the Twitter lists API.[10] For example, the Qatari minister of foreign affairs[11] is a member of *Foreign Ministers* Twitter list[12], which lists foreign ministers of other countries.[13]

To further ensure that the users in our collection tweet in Arabic, we only kept the users that tweeted at least one single Arabic tweet out of their last 10 tweets. Our final user collection consists of 395,231 users, where 9,261 of them (2.34%) are verified.

*3.4.2.3. Users Social Data*

For each user in our collection, we collected the following:

- **Timeline:** We retrieved the full timeline (max 3,200 as allowed by Twitter user_timeline API).[14] In total, 877,470,516 Arabic tweets were collected. User timelines can be used to extract signals about the users' expertise and topics of interest.

- **Twitter lists:** Using the Twitter lists API, we retrieved all the Twitter lists the user is a member of, as lists are shown to be effective in expert finding in social media studies [139], [143]. We end up with 7,353,520 Twitter lists, where 1,192,284 are unique.

Table 3.1 gives an overall summary of our user collection. An example of an authority along with his collected profile metadata, some of his Twitter lists, and an example of an Arabic tweet collected from his timeline is presented in Figure 3.2.

### *3.4.3. Human Annotations*

We hired two graduate annotators to find Twitter authority accounts (the maximum they can find) for each rumor in our collection. We asked the annotators to follow the below guidelines during their annotation process:

[9] https://developer.twitter.com/en/docs/twitter-api/users/follows/api-reference/get-users-id-following

[10] https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/create-manage-lists/api-reference/get-lists-list

[11] https://twitter.com/MBA_AlThani_

[12] https://twitter.com/i/lists/1547472831914643457

[13] https://twitter.com/i/lists/1547472831914643457/members

[14] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user_timeline


Figure 3.2. An authority with his profile, some of his Twitter lists, and an Arabic tweet from his collected timeline.

Table 3.1. Summary of the user collection.

| | |
|---|---|
| Users | 395,231 |
| Verified | 9261 (2.34%) |
| Listed in Twitter lists | 291,653 (74%) |
| Timeline tweets | 877,470,516 |
| Average number of collected tweets per user | 2220.15 |
| Unique Twitter lists | 1,192,284 |
| Average number of lists per user | 18.6 |

1. **Authorities from fact-checkers:** Given a rumor and its corresponding fact-checking article, the annotators study first the authorities mentioned or contacted by the fact-checkers to verify the rumor, to check if they have Arabic Twitter accounts. Moreover, the annotators check if the sources of evidence provided by the fact-checkers, e.g., Web pages or other social media accounts such as Facebook, have equivalent Twitter accounts.

2. **Expanding the authority list:** The annotators then expand the authority list above with the help of Twitter and Google search services. Specifically, they start with the authorities found above and also the entities mentioned in the rumor, and explore more related entities, if they are relevant to the rumor, as follows:

   (a) *Organizations:* The annotators search for the organization's Twitter account, its president, secretary, spokesman, media channel, etc.

   (b) *Persons:* The annotators look for the person's Twitter account, organization, spokesman, etc.

Table 3.2. Summary of rumor collection and relevance judgments statistics.

| | **Political** | [min, max] | **Sports** | [min, max] | **Health** | [min, max] | **Overall** | [min, max] |
|---|---|---|---|---|---|---|---|---|
| **Rumors** | 50 | | 50 | | 50 | | 150 | |
| **Authorities** | 406 | [2, 18] | 246 | [1, 14] | 392 | [1, 22] | 1044 | [1, 22] |
| *Relevant* | 158 | [0, 10] | 77 | [0, 6] | 190 | [0, 13] | 425 | [0, 13] |
| *Highly Relevant* | 248 | [0, 17] | 169 | [0, 13] | 202 | [0, 11] | 619 | [0, 17] |

(c) *Countries:* The annotators check the country's president, ministers, ministries, embassies, ambassadors, spokesman, government media channel, etc.

For some rumors, searching the Web is needed to know the leaders, organizations, spokesmen, etc., for specific domains in specific countries. For example, if the rumor is about a health issue in Egypt, then the annotators can search for the name of the health minister,[15] ministry,[16] or other leaders in health organizations in Egypt, such as WHO.[17]

It is worth noting that we asked annotators to consider news agencies or journalists as authorities ***only if*** the rumor is about them, or the entities mentioned in the rumor are related to them. Finally, the annotators considered only the *active* Arabic Twitter accounts, defined as those who posted at least one tweet around the time of the rumor (one month before and one month after).

### *3.4.4. Graded Relevance*

Given that some authorities might be highly relevant while others are not, we asked the annotators to indicate whether they think the authority is *highly relevant* or *relevant* to the rumor, defined as follows:

- **Highly relevant (Grade 2):** if the authority should be contacted first (or has a high priority to be contacted) to help verify the rumor.
- **Relevant (Grade 1):** if the authority has a lower priority to be contacted.

For example, the rumor stating *"Foreign Minister Najla Al-Manqoush: Turkey is seeking to ignite the war again and disrupt the elections."*, the foreign ministry[18] and minister of Libya[19] are examples of *highly relevant* authorities, while the president[20] and spokesman[21] of the high council of state of Libya are considered just *relevant*. Table 3.2 gives overall statistics of the annotations with minimum and maximum number of authorities for the rumors.

[15] https://twitter.com/DrHalaZayed1
[16] https://twitter.com/mohpegypt
[17] https://twitter.com/WHOEgypt
[18] https://twitter.com/Mofa_Libya
[19] https://twitter.com/NajlaElmangoush
[20] https://twitter.com/KhaledMeshri
[21] https://twitter.com/NajWheba

### *3.4.5. Data Quality*

We assigned two annotators for each rumor, and each separately may find different or additional authorities than the other. Therefore, we asked them to meet afterward and discuss the cases they missed or disagreed with. For the remaining cases of disagreements on authorities or their graded relevance after the meeting, a third annotator was hired to resolve them. Specifically, 3.88% and 2.16% of the annotated authorities and graded relevance, respectively, were resolved by a third annotator. We found that the overlap between the two annotators was on 36% and 90% of the authorities before and after their meeting, respectively.

To measure the quality of our data, we considered all authorities or graded relevance that both annotators agreed upon as cases of agreement; otherwise, they were considered cases of disagreement. Based on that, the computed Cohen's Kappa inter-annotator agreement [158] was found to be 0.78 and 0.71 on authorities and graded relevance annotations, respectively, which indicates a *substantial* agreement [159] for both.

### *3.4.6. AuFIN vs. Misinformation Expert Finding Datasets*

In Table 3.3, we demonstrate a comparison between our data and the aforementioned misinformation expert finding test collection[136], [137] (refer to Section 3.2).

Table 3.3. Comparison between AuFIN and existing misinformation expert finding test collections.

| Work | Platform | Lang | #Users | #Rumors | #Experts/ Authorities | Annotated/ Released |
|---|---|---|---|---|---|---|
| [136] | Weibo | Chinese | 5M | 859 | | |
| [137] | Twitter | English | 491,622 | 20 | | |
| **AuFIN** | Twitter | Arabic | 395,231 | 150 | 1044 | ✓ |

## 3.5. Proposed Approach

In this section, we present our proposed model to tackle the problem of authority finding. Section 3.5.1 details the basic architecture of our model. Section 3.5.2 presents our proposed approach for rumor expansion to further improve the performance.

### *3.5.1. Authority Finding Model*

Figure 3.3 depicts an overview of our authority finding model, which retrieves a ranked list of authorities, represented by their Twitter accounts, given a rumor expressed in a tweet. The model employs an ad-hoc retrieval approach for initial retrieval, and pre-trained language models for semantic reranking [69]. It is mainly composed of three stages. The first is *Initial Retrieval*, which adopts an unsupervised approach to retrieve a set of candidate users using *lexical* matching with the textual content of their profiles, in addition to incorporating Twitter-based social/network features to estimate their authoritativeness. The second is *Semantic Reranking*, which adopts a supervised approach that exploits contextualized pre-trained language models to rerank the users

retrieved by the first stage using a textual representation of their profiles. The third is *Hybrid Reranking*, which combines the output of the first two stages to rerank the initially-retrieved candidate users. The following subsections describe the three stages in detail.

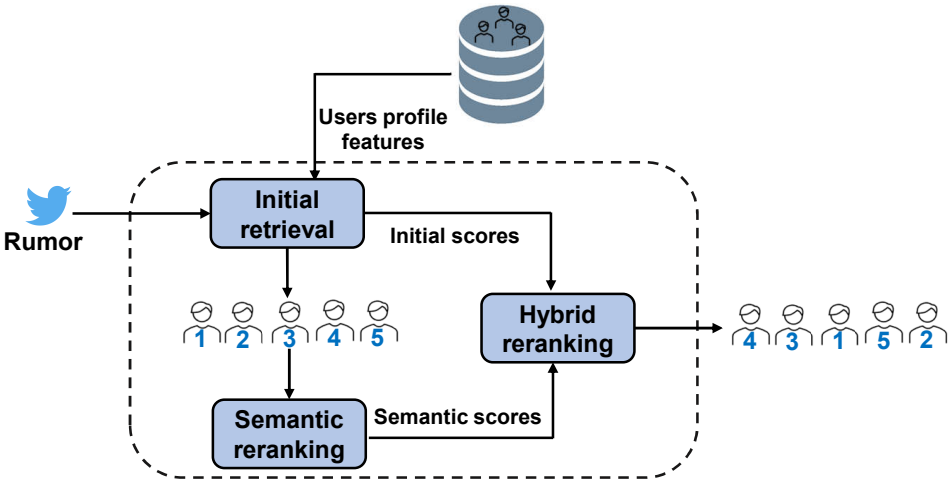



Figure 3.3. Overview of our authority finder model.

### *3.5.1.1. Initial Retrieval*

For *initial retrieval*, the similarity between rumor and user's textual representation is first computed by lexical matching using traditional retrieval models. A major question here is how we can create a textual representation for each user, using her profile, that is relevant to our problem. Ghosh, Sharma, Benevenuto, *et al.* [139] showed that Twitter *lists* constitute a valuable "crowd-sourced" resource for topic expert finding in Twitter, and subsequent studies for topic and misinformation expert finding in Twitter further confirmed those findings [137], [142], [143]. In our work, we experimented with different user's textual representations using a concatenation of the name and description of the Twitter lists she is listed in, in addition to her profile's name and description, and timeline tweets.

To compute the lexical score $S_x$ for a candidate user $u$ given a rumor $r$, we adopt BM25 retrieval model [160], one of the most successful term-weighting techniques widely adopted as a starting point from many text ranking methods by researchers and commercial systems [69], [161]–[163].

$$S_x(u, r) = \text{BM25}(t_u, t_r), \tag{3.1}$$

where $t_u$ and $t_r$ are the textual representations of the user $u$ and rumor $r$ respectively.

The lexical score considers only the textual representation of users, but it neglects other user profile's features that can constitute evidence of authoritativeness. This has also been explored in prior studies. Liang, Liu, and Sun [136] and Li, Dong, Yang, *et al.* [137] assumed that experts usually have more followers than regular social media users; hence they considered the (logarithm of) followers count divided by followees count in scoring users as experts. Ghosh, Sharma, Benevenuto, *et al.* [139] assumed that the more Twitter lists a user is listed in, the more likely she is an expert. In our work, we

adopt/combine both assumptions, where we assume that authorities have more followers *and* are involved in more Twitter lists than regular users. Therefore, we compute the *initial score* $S_i(u, r)$ of a candidate user $u$ given a rumor $r$ as follows:

$$S_i(u, r) = S_x(u, r) \times log_2[(l_u + 2)(\frac{f_u}{w_u} + 2)], \tag{3.2}$$

where $l_u$ is the number of Twitter lists the candidate is a member of, $f_u$ is the number of followers the candidate has, and $w_u$ is the number of users the candidate is following. It is worth noting that our initial scoring function takes into consideration that some authorities may not be listed in any Twitter lists, and in this case, the followers and followees count will be a measure of their authoritativeness. Note that we add 2 when computing the logarithm of both factors for smoothing in case $l_u$ or $f_u/w_u$ is zero. In any of those cases, the initial score will fall back to the *lexical* score.

#### *3.5.1.2. Semantic Reranking*

Contextualized transformer-based models such as BERT [95] that are pre-trained on large corpora have shown superiority in document reranking, e.g., [164]–[166]. Variant pre-trained models on Arabic corpora were released recently, e.g., [167]–[169], and they achieved SOTA results in different Natural Language Processing (NLP) and Information Retrieval (IR) tasks in general [170]–[173] and document ranking in particular [48], [174].

While our initial score captures solely the lexical matching between the rumor and the user textual representation, pre-trained language models can capture the *semantic* matching to mitigate the vocabulary mismatch issue, i.e., the rumor terms do not have to lexically match the user representation terms in order to contribute to relevance. We adopt monoBERT [164], which was shown to outperform neural retrieval models [69] and is widely used recently by the IR community for text reranking [174]–[176], to fine-tune the pre-trained models for user authority relevance. We feed BERT the tweet containing the rumor as sentence $A$, and the user's textual representation as sentence $B$ separated by the [SEP] token. Finally, we use the contextual representation of the [CLS] token ($c_{cls}$) as input to a single classification layer, with two output nodes, added on top of the BERT architecture to compute the probability of the user $u$ being authority for the rumor $r$, as follows:

$$S_s(u, r) = P(Relevance = 1|t_u, t_r) = \text{softmax}(c_{cls}(t_u, t_r) \times W + b)_1, \tag{3.3}$$

where $c_{cls} \in \mathbb{R}^D$, $W \in \mathbb{R}^{DX2}$, $D$ is the embedding dimension of the model, and $b \in \mathbb{R}^2$ is the bias vector for the 2 output classes. $softmax(.)_1$ indicates the probability of the *relevant* class. Following Nogueira and Cho [164], we train the model for relevance classification using cross-entropy loss:

$$Loss(r) = -\sum_{u \in U_+} log(S_s(u, r)) - \sum_{u \in U_-} log(1 - S_s(u, r)), \tag{3.4}$$

where $U_+$ and $U_-$ are the set of positive and negative candidate users respectively, which are basically the training data used to fine-tune the model. Similar to the work

by Mansour, Elsayed, and Al-Ali [174], we generate the training examples by choosing the positive examples from the annotated data and negative examples randomly from the top initially-retrieved candidates.

#### *3.5.1.3. Hybrid Reranking*

Given that the *initial* score incorporates the user profile features, i.e., Twitter lists count, followers, and followees count, which was shown to be crucial in measuring the popularity of users, we propose a *hybrid* approach that combines both the *initial* and the *semantic* scores to compute a final score of candidate users, as follows:

$$S_h(u, r) = \alpha \times \hat{S}_i(u, r) + (1 - \alpha) \times S_s(u, r) \tag{3.5}$$

where $\alpha \in [0, 1]$ is a weight that indicates the relative importance of each score, and $\hat{S}_i(u, r)$ is the normalized initial score using min-max normalization per rumor.

### *3.5.2. Rumor Expansion*

In our work, we consider tweets as the source of rumors. However, due to the short length of tweets and the entity sparsity issue [177], where in many cases a tweet mentions only one entity, the retrieval of authorities by textual matching can be challenging. To mitigate this issue, we propose expanding the rumor with a longer textual representation of the mentioned entities and other related entities by exploiting Knowledge Bases.

A *Knowledge Base* (KB) is "a repository of entities with information about their relationships and attributes in a (semi-) structured format" [178], e.g., Wikipedia. A *Knowledge Graph* (KG) is "a knowledge base that is specifically represented as a graph, where entities, attributes, and relations are represented through nodes and edges in the graph." [178], e.g., DBpedia [179], WikiData [180], and Freebase [181].

Our hypothesis is that entity expansion can help the retrieval of relevant authorities who are not explicitly mentioned in the rumor, in addition to the disambiguation of those who are already mentioned. As presented in Figure 3.4, we propose the following techniques to expand the set of entities mentioned in the rumor:

1. **KB-based:** Our objective here is to link the mentioned entities in the rumor to their associated entities in the KB, find related entities to them, then expand the rumor with a (hopefully indicative) textual representation of both the linked and expanded entities. To our knowledge, no prior studies have examined entity expansion by exploiting the KB for expert finding in social media. Our expansion approach, presented in Figure 3.4, has the following steps:

   (a) **Named Entity Recognition (NER):** We first extract the mentioned entities using NER to identify the named entities, namely, persons, locations, or organizations, mentioned in the rumor. We believe they are the key elements to retrieving relevant authorities.

   (b) **Entity Linking (EL):** We then associate each extracted mentioned entity with an entity in the KB [182]. Usually, named entities could have variant surface forms, such as their full names, partial names, aliases, and abbreviations. We perform this step to disambiguate the extracted mentioned entities.

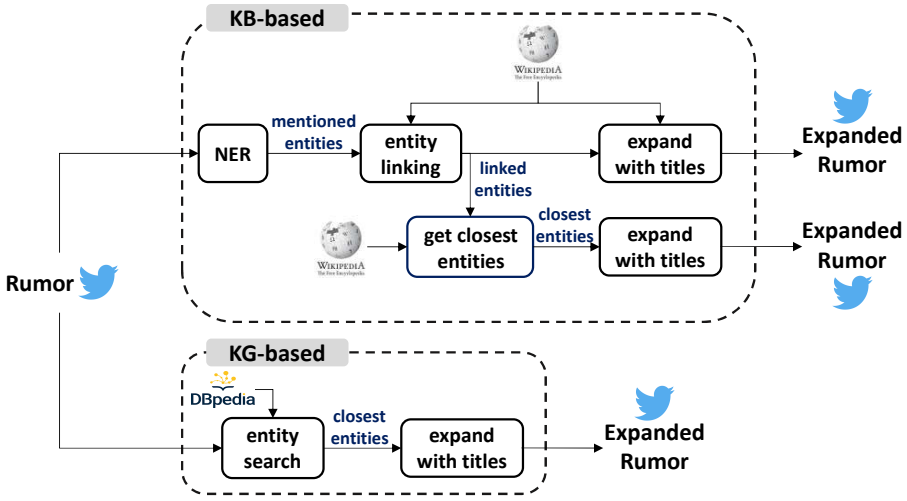



Figure 3.4. Our approaches for rumor expansion.

(c) **Rumor Expansion:** Using the linked entity, we adopt two approaches to expand the rumor:

- *Exploiting entity's textual representation:* Each entity in a KB has a textual representation, e.g., title, abstract, categories, etc. Motivated by the fact that the entities mentioned in the tweets are usually ambiguous due to length limitations, in our work, we consider the title of the entity in the KB as its textual representation and use it to expand the rumor. For example, the rumor *"Erdogan: My problems with Al-Sisi were due to a misunderstanding, and I seek to fix it. Al Jazeera"* mentions two persons, namely *Erdogan* and *Al-Sisi*, which we believe, if disambiguated/expanded, they can help in retrieving relevant authorities. These mentions were then expanded to *Recep Tayyip Erdoğan* and *Abdel Fattah El-Sisi* respectively, using their linked entities in the KB.
- *Exploiting the entity embeddings:* Entity embeddings can provide richer vector representation of entities in the KB [183]. In our work, we utilize entity embeddings to find the most similar entities in the KB for each mentioned entity in the given rumor, using cosine similarity. We then use the top N most similar entities to expand the rumor with their textual representation. For the example rumor mentioned earlier, the nearest entities to *Erdogan* and *Al-Sisi* are *President of Turkey* and *President of Egypt*, respectively. Given that the tweet does not mention Turkey or Egypt, we believe expanding the tweet with these keywords can improve the retrieval of relevant authorities.

2. **KG-based:** We can also find relevant entities by free-text *entity search*. To achieve that, we adopted the entity search task [184] where we consider the given rumor tweet as the query to KEWER system proposed by Nikolaev and Kotov [185] for the task. As far as we know, entity search was addressed as a standalone task, but it was not adopted for query expansion in general, and for rumor context expansion in particular.

It is worth noting that we use the expanded query for initial retrieval but use the

original query for reranking, inspired by previous studies that showed that expanding the query without reformulating it in natural language may not be useful for BERT [69].

## 3.6. Experimental Setup

For the reproducibility of our results, we provide here some implementation details of our work at all phases, in addition to evaluation setup for our experiments.

### *3.6.1. Initial Retrieval*

- *Preprocessing*: Twitter users usually use different languages for their username and profile description or even combine two languages, such as Arabic and English. Moreover, most of the Twitter list names and descriptions are in non-Arabic text. Given that our objective is to find Twitter accounts of authorities for Arabic rumors, and textual content is one of the crucial components in our system, we translated the non-Arabic text in the name and description of user profiles and Twitter lists in our test collection using Google Translate API,[22] and concatenated the translation with the Arabic text if it exists.

- *Constructing User Documents*: We constructed 7 variant text representations of users (denoted by *documents* for the purpose of indexing) in order to perform an ablation study.[23] Specifically, we constructed *bio* (profile name and description), *lists* (name and description of all Twitter lists of which the user is a member, in the order retrieved by Twitter), *timeline* (the user's collected timeline tweets, in the reverse chronological order), and combinations of them.

- *Indexing and Retrieval*: To index our user collection for *lexical* retrieval, we used Pyserini [186], a Python toolkit for information retrieval. For BM25, we used default values of parameters ($k_1 = 0.9$ and $b = 0.4$). For indexing and retrieval, we set the language to Arabic. Pyserini adopts Lucene Arabic analyzer[24] for stemming and removing stop words. We performed normalization on both documents and queries.[25] Given that our queries (i.e., rumors) are expressed in tweets, we discarded URLs, emojis, and non-Arabic characters as preprocessing.

### *3.6.2. Semantic Reranking*

We experiment with fine-tuning 7 variant Arabic pre-trained language models, selected based on the corpora used for their pre-training. Specifically, we selected those that covered news articles, Arabic Wikipedia dumps, Arabic Common Crawl, or Arabic tweets. The adopted models are AraBERT [167], Arabic BERT [187], GigaBERT [188], MARBERT [168], ARBERT [168], CAMeLBERT-Mix [169], and AraELECTRA [189]. All models were pre-trained by adopting the BERT architecture [95]; an exception is AraELECTRA which is based on ELECTRA architecture [190]. In our experiments, we considered the base version (constituting 12 encoder layers) of all

[22] `https://cloud.google.com/translate`

[23] We share our indexed data for reproducibility.

[24] `https://lucene.apache.org/core/4_6_0/analyzers-common/org/apache/lucene/analysis/ar/ArabicAnalyzer.html`

[25] We used the normalization code available at `https://alraqmiyyat.github.io/2013/01-02.html`.

BERT and ELECTRA models for consistency and fair comparison across models. We reranked the top 100 initially-retrieved users, denoted as *candidate users*. For developing our models, we utilized Hugging Face transformers library [191] v4.11.1 and PyTorch 1.7.1 with CUDA 11.0.

### *3.6.3. Hyper-parameter Tuning*

We tuned our models using 5-fold cross-validation, optimizing for P@5 on each fold. The 5 folds were split such that each fold contains 30 rumors, 10 from each category of rumors,[26] i.e., political, health, and sports. Specifically, we tuned the following hyper-parameters within the following values: learning rate [2e-5, 3e-5], dropout [0.2, 0.3, 0.4], epochs [3, 4, 5], and $\alpha$ [between 0 and 1 with 0.05 increments]. The batch size, max sequence length, and reranking depth were fixed to 16, 512, and 100 respectively. AdamW [192] was used to optimize our models. In our experiments, we used 3 folds for training and fine-tuned on a fourth fold for validation. Once we found the best values of the hyperparameters using the validation fold, we trained our models on 4 folds (training + validation folds) and tested on the remaining fifth fold. We trained all our models on a Linux machine running Intel Xeon E5-2690 v4 CPU, Tesla P100 16GB GPU, and 128GB RAM. The training and validation per epoch took an average of 1 min and 50 seconds.

### *3.6.4. Rumor Entity Expansion*

- **KB-based:** We used CAMeL NER model [193] for Arabic named entity recognition[27] to extract entities from each tweet. CAMeL is a BERT-based NER model that was shown to outperform existing Arabic NER models [193]–[195]. For entity linking and retrieving the textual representation, we adopted the Wikipedia API.[28] We then expanded the tweet by exploiting the Wikipedia page of the extracted entity as follows:
  - *Entity title:* For each entity mentioned in the tweet, specifically persons, organizations, or locations, we used the title of the linked Wikipedia page for expansion.
  - *Entity embeddings:* We trained the Wikipedia2vec model [183] using Arabic Wikipedia dumps of May 2022. For each entity extracted from our tweet, we retrieved the closest top 5 entities in the embedding space using our model[29] and used the Wikipedia title of the entities for rumor expansion.
- **KG-based:** We adopted KEWER [185], a SOTA entity search model at the time of performing the experiments, as a model to search for relevant entities for our rumors from KG. The proposed model, targeting English entity search, is composed of three stages. The first is initial retrieval using BM25 from a DBpedia pages index, to retrieve an initial potentially-relevant DBpedia entities. The second estimates the relevance of the retrieved entities measured by the

[26] We release the the 5 folds for the reproducibility of the experiments.

[27] `https://camel-tools.readthedocs.io/en/latest/api/ner.html`

[28] `https://github.com/goldsmith/Wikipedia`

[29] We release the Arabic trained Wikipedia2vec model to the research community for the reproducibility of the experiments.

cosine similarity between the query and the embeddings of the entities. The embeddings are extracted using a word2vec model trained by the authors using DBpedia pages. Finally, the lexical and the similarity scores are interpolated to rerank the initial DBpedia entities using the hybrid score. Given that 98% of our rumors were propagated on Twitter in the period of 2020 and 2021, we used the Arabic DBpedia 2021-11[30] (the latest at that time) to train the word2vec model and to have a collection of DBpedia pages for initial retrieval. We used Pyserini to index and retrieve the initial candidates. To rerank the retrieved candidates, we trained the word2vec model using the authors' code.[31] With the lack of labeled data for this specific entity search task, i.e., (rumor, DBpedia entity), we could not fine-tune the model, so we used the hyper-parameter values adopted by Esmeir [196], who used KEWER for Arabic entity search. We used the tweet expressing the rumor as a query to search for entities, and used the top 5 retrieved entities' DBpedia titles to expand the rumor.[32]

#### *3.6.5. Evaluation Measures*

To measure the effectiveness of our proposed model over different setups, we compute precision at ranks 1 and 5 (P@1 and P@5), Normalized Discounted Cumulative Gain at rank 5 (NDCG@5), and Recall at rank 100 (R@100), which are widely used measures to evaluate the performance of text retrieval systems [69], [197]. P@1 and P@5 show how our model is able to retrieve authorities at the top of a *short* retrieved list of depths 1 and 5, supporting the scenario of a journalist (or normal Twitter user) seeing a viral rumor and wanting to check a short list of trusted sources. NDCG@5 is also selected to measure how the model is able to retrieve highly relevant authorities up in the list. P@k and NDCG@k were also adopted by related misinformation expert finding studies [136], [137]. We also measure the recall at depth 100 (and sometimes at depths 20 and 50) for the purpose of studying how some setups are improving the recall of authorities.

Moreover, we report statistical significance using the two-sided t-test with a significance level of 5% on all evaluation metrics. Furthermore, we applied Benjamini-Hochberg correction [198] to adjust $p$-values to control the false discovery rate.

#### *3.6.6. Baselines*

We compare our proposed model with SOTA models. We first considered the only two misinformation-based expert finding models mentioned earlier [136], [137]. We were not able to implement the model proposed by Liang, Liu, and Sun [136], because they proposed training a tag suggestion model and an entity linking to micro-blog user model, which both need training data that is not available in Arabic. The model proposed by Li, Dong, Yang, *et al.* [137], on the other hand, was not implemented due lack of needed details.[33] We, then, were left with models for topic expert finding in social media; we selected two SOTA models, a profile-based (Cognos) and graph-based

[30] `https://databus.dbpedia.org/dbpedia/`

[31] We modified the preprocessing step to adapt it to Arabic text.

[32] We release the Arabic trained word2vec model and the DBpedia index for the reproducibility of the experiments.

[33] We contacted the authors for clarification with no response.

(FAME) models, in addition to a document-based approach proposed for expert search (CombMNZ).

- **Cognos** [139]: A topic expert finding profile-based model that leverages Twitter lists solely to rank experts relevant to a query. The authors constructed a topic vector for each user using the name and description of their Twitter lists, then computed the similarity between the query and user vectors. Finally, the users were reranked by multiplying the similarity by the logarithm of their list count. The authors adopted the cover density algorithm [199] for one to three-term queries. Given that our queries are much longer, we adopted the cosine similarity on TF-IDF based representations, per authors' suggestion. We considered only users who are members of at least 10 Twitter lists, as proposed by the authors.

- **FAME** [143]: A topic expert finding graph-based model that leverages Twitter lists and adopts the personalized PageRank algorithm to find "important" nodes (users) in the edge-labeled graph given a query. In the original method, the authors started the graph construction with a seed of "expert" users in three topics. Through an iterative process, they collected Twitter lists for users in the graph, where the user who created the list is considered an endorser. At the end of each iteration, they prune the graph to exclude users with the lowest in-degree or out-degree. In our work, we constructed the graph with all users in our user collections and pruned them to exclude the ones that have an in-degree or out-degree of 5 or less. It is worth mentioning that FAME's underlying graph was only constructed for three topics, while our graph was constructed to include all users in our collection who have variant backgrounds and expertise.

- **CombMNZ** [200]: A document-based voting model proposed for the expert search task defined in TREC Enterprise track 2005 [201]. The model first retrieves documents relevant to the query, then scores each user by the relevance scores of their retrieved documents. To implement this method in our context, we indexed all individual tweets (considered as individual documents) in our collection using Pyserini [186]. We retrieved the top 1,000 potentially relevant tweets for each query, then computed the user scores per the proposed model.

## 3.7. Experimental Evaluation

In this section, we present and discuss the results of our experiments that address the following research questions:

1. **RQ1:** What is the best user textual representation for lexical retrieval? Will incorporating the network features improve the performance? (Section 3.7.1)

2. **RQ2:** Will semantically-reranking the initially-retrieved users improve the performance? And what is the effect of combining both the semantic and initial retrieval scores? (Section 3.7.2)

3. **RQ3:** Will exploiting the KB for rumor expansion further improve the retrieval of authorities? (Section 3.7.3)

4. **RQ4:** How do existing topic expert finding models perform compared to our models? (Section 3.7.4)

### 3.7.1. User Representation and Network Features for Initial Retrieval (RQ1)

To address RQ1, we retrieve the top 1,000 candidate users for each rumor, using seven different user textual representations mentioned in Section 3.6. Table 3.4 shows the performance of both *lexical* retrieval, using the variant user representations, and the *initial* retrieval, which incorporates the user network features.

Table 3.4. Lexical and initial retrieval with variant user representations. A star indicates statistically significant improvement of initial model over the lexical *lists* model. Bold and underlined numbers indicate the best and second-best performance in each retrieval type per evaluation measure.

| Retrieval | Representation | P@1 | P@5 | NDCG@5 | R@100 |
|---|---|---|---|---|---|
| Lexical | *bio* | 0.100 | 0.049 | 0.060 | 0.153 |
| | *lists* | **0.193** | **0.103** | **0.141** | **0.329** |
| | *timeline* | 0.007 | 0.003 | 0.003 | 0.021 |
| | *bio + lists* | <u>0.180</u> | <u>0.097</u> | <u>0.119</u> | <u>0.318</u> |
| | *bio + timeline* | 0.007 | 0.003 | 0.003 | 0.021 |
| | *lists + timeline* | 0.007 | 0.003 | 0.003 | 0.024 |
| | *bio + lists + timeline* | 0.007 | 0.003 | 0.003 | 0.024 |
| Initial | *bio* | 0.193 | 0.095 | 0.134 | 0.238 |
| | *lists* | <u>0.347</u>$^\star$ | <u>0.151</u>$^\star$ | <u>0.221</u>$^\star$ | **0.428**$^\star$ |
| | *timeline* | 0.027 | 0.017 | 0.019 | 0.081 |
| | *bio + lists* | **0.373**$^\star$ | **0.167**$^\star$ | **0.238**$^\star$ | <u>0.425</u>$^\star$ |
| | *bio + timeline* | 0.027 | 0.019 | 0.020 | 0.081 |
| | *lists + timeline* | 0.027 | 0.020 | 0.023 | 0.090 |
| | *bio + lists + timeline* | 0.027 | 0.021 | 0.024 | 0.090 |

#### 3.7.1.1. ***Lexical Retrieval***

The results of the *lexical* retrieval show clearly that relying on the user profile name and description as a user representation (denoted as *bio*) is insufficient. In the contrary, exploiting the user's *lists* is crucial and can improve the performance significantly. However, using the user's *timeline* severely degraded the performance, even when it is combined with *lists*; we investigate why in our failure analysis in Section 3.8.1. The results then demonstrate that *lists* representation exhibits the best performance of individual textual user representation for lexical retrieval in all evaluation measures.

#### 3.7.1.2. ***Initial Retrieval***

The results of the initial retrieval yield several interesting observations. First, performance improved with all representations in all evaluation measures, showing a clear and very strong positive impact of using user network features. Second, combined *bio + lists* user representation for *initial* retrieval yielded the best retrieval performance with a difference over *lists* that is statistically significant in both P@5 and NDCG@5. Furthermore, *bio+lists* improved R@100 by 33.65% compared to *lexical* retrieval. While *lists* representation exhibits the best R@100 performance, *list+bio* has a comparable performance that is statistically insignificant from *lists*. As we are aiming to rerank

the candidates from initial retrieval and also combine evidence from both lexical and semantic retrieval, both recall and precision are important; therefore, we adopt *bio+lists* representation in all subsequent experiments.

To answer **RQ1**, we conclude that, first, the timeline user representation is inadequate for the retrieval of authorities, even when combined with other textual representations. Second, the user's bio and subscribed lists combined are the most effective representation among the ones we experimented with. Finally, user network features contribute significantly to the authority finding performance.

### *3.7.2. Semantic Reranking of Candidates and Combining Signals (RQ2)*

To address RQ2, we fine-tuned seven Arabic BERT-based pre-trained language models to classify whether a user is an authority for a given rumor or not, and then reranked the top 100 candidate users (that were initially retrieved) using the relevance scores produced by the fine-tuned models. We denote this as "semantic reranking" due to the use of pre-trained models. We feed the BERT model the rumor expressed in the tweet and the user representation, structured as follows: *"[CLS] rumor [SEP] user representation [SEP]"*, where, for each candidate, we used the *bio + lists* representation, as mentioned in Section 3.7.1. Given that the input to BERT is limited to 512 tokens while the user representation may exceed that length, we truncated the representation whenever needed.

Table 3.5 presents the results of reranking the top 100 candidate users retrieved initially using *bio+lists* representation. Focusing first on precision-based measures, none of the semantic models outperforms the initial retrieval baseline, i.e., the semantic signal solely is less effective in terms of precision than lexical and network signals. We also notice that there is no clear winner among the semantic models.

However, all our precision-based measures consider only the top 5 candidates, but are semantic models able to position some good candidates a bit lower (or not very far) in the list? To check that, we also measure R@20 and @R@50 here. Surprisingly, several semantic models outperform the baseline, yielding statistically significant improvements of up to 13.58% on R@50. This shows that the semantic signal has potential in our task, and that combining it with the lexical and network signals is promising.

That encouraged us to conduct the subsequent experiment, in which we combined both the initial and semantic scores by linear interpolation using Equation 3.5.1.3. We then reranked the users based on the resulted *hybrid* scores. Table 3.6 presents the performance of the models after interpolation. We note that both *Arabic BERT* and *MARBERT* models yield improved performance over the baseline on almost all measures with statistically significant differences on P@5, R@20, and R@50. The tuned interpolation factor $\alpha$ ranges between 0.75 and 0.77, which indicates that the initial retrieval score is more indicative than the semantic score in our setup, but both still complement each other.

Table 3.5. Reranking top 100 candidate users initially-retrieved by Initial (*bio+lists*). A star indicates statistically significant difference compared to the initial retrieval baseline. Bold and underlined numbers indicate the best and second-best performance per evaluation measure.

| **Model** | **P@1** | **P@5** | **NDCG@5** | **R@20** | **R@50** |
|---|---|---|---|---|---|
| Initial (*bio+lists*) | **0.373** | **0.167** | **0.238** | 0.263 | 0.346 |
| *AraBERT* | 0.147 | 0.133 | 0.115 | 0.273 | 0.381$^\star$ |
| *GigaBERT* | 0.147 | <u>0.139</u> | 0.133 | <u>0.289</u> | **0.393**$^\star$ |
| *Arabic BERT* | 0.160 | 0.128 | 0.128 | **0.300** | <u>0.390</u>$^\star$ |
| *MARBERT* | 0.160 | 0.107 | 0.113 | 0.232 | 0.318 |
| *ARBERT* | 0.120 | 0.113 | 0.114 | 0.279 | 0.379$^\star$ |
| *CAMeLBERT-Mix* | 0.160 | 0.136 | <u>0.138</u> | 0.279 | **0.393**$^\star$ |
| *AraELECTRA* | <u>0.180</u> | 0.127 | 0.137 | 0.267 | 0.367 |

We also note that *ARABIC BERT model* exhibits the best performance over all measures (except R@50 with a slight difference). *Arabic BERT* is the only model, among the ones we experimented with, that was trained only on Wikipedia dumps and the unshuffled Oscar data.[34] We notice that Wikipedia page titles are quite similar in content and writing style to the Twitter list names and descriptions, which are the main input to the model in our experiments; that may explain why it achieved the best overall performance.

It is also worth mentioning that despite being fine-tuned only using binary relevance, (almost) all integrated models outperformed the initial retrieval baseline on NDCG@5, which considers graded relevance.

Table 3.6. Reranking top 100 candidate users initially-retrieved users by initial (*bio+lists*) by **interpolating** initial and semantic scores. A star indicates statistically significant difference compared to the initial retrieval baseline. Bold and underlined numbers indicate the best and second-best performance per evaluation measure.

| **Model** | **P@1** | **P@5** | **NDCG@5** | **R@20** | **R@50** |
|---|---|---|---|---|---|
| Initial (*bio+lists*) | <u>0.373</u> | 0.167 | 0.238 | 0.263 | 0.346 |
| *+AraBERT* | 0.327 | 0.181 | 0.235 | 0.301$^\star$ | 0.388$^\star$ |
| *+GigaBERT* | 0.327 | 0.189 | 0.238 | 0.317$^\star$ | <u>0.398</u>$^\star$ |
| *+Arabic BERT* | **0.413** | **0.213**$^\star$ | **0.271** | **0.333**$^\star$ | 0.396$^\star$ |
| *+MARBERT* | 0.367 | 0.184$^\star$ | 0.246 | 0.297$^\star$ | 0.378$^\star$ |
| *+ARBERT* | 0.367 | 0.175 | 0.237 | 0.291$^\star$ | 0.384$^\star$ |
| *+CAMeLBERT-Mix* | 0.333 | <u>0.199</u> | 0.239 | <u>0.319</u>$^\star$ | **0.403**$^\star$ |
| *+AraELECTRA* | 0.360 | 0.193 | <u>0.255</u> | 0.315$^\star$ | 0.397$^\star$ |

Table 3.7 presents a comparison between the top 5 retrieved users using the *initial* and the *hybrid* models, where *bio+lists* user representation and *Arabic BERT* were adopted. For the (political) rumor, the *initial* retrieval alone was not able to

[34] `https://oscar-corpus.com/`

retrieve any authority within the top 5, whereas the *hybrid* model managed to retrieve three *highly relevant* authorities at ranks 2, 4, and 5.

To answer **RQ2**, we conclude that *semantic* ranking is inferior to the lexical and network-based ranking in terms of precision but superior in terms of recall. That led to a hybrid model that is better in both precision and recall.

Table 3.7. Top 5 retrieved authorities by *initial* and *hybrid* retrieval. Non-underlined user names are *non-relevant* and double underlined ones are *highly relevant*.

| **Rumor** Watch .. How #Qataris celebrated in the streets of Doha after the Kingdom agreed to open the land and air borders with their country. | | | |
|---|---|---|---|
| **Initial Retrieval** | | **Hybrid Retrieval** | |
| 1. @qatarweather | Qatar Meteorology Department - Civil Aviation Authority. | 1. @qatarweather | Qatar Meteorology Department - Civil Aviation Authority. |
| 2. @kataraqatar | Katara, is one of the grand and unique cultural projects worldwide. | 2. @MOI_Qatar | Learn about the Ministry's activities and services through its official account. Ministry of Interior Affairs, Qatar |
| 3. @VodafoneQatar | Official Twitter page of Vodafone Qatar. | 3. @roadto2022 | Official account for Qatar's FIFA World Cup 2022™ delivery and legacy organisation. |
| 4. @ClimatJazeera | to be in the event... | 4. @MofaQatar_AR | The official account of the Qatari Ministry of Foreign Affairs. |
| 5. @OoredooQatar | The virtual place for the Ooredoo team to find out about the latest offers. | 5. @MBA_AlThani_ | Deputy Prime Minister & Minister of Foreign Affairs, Qatar. |

### *3.7.3. Exploiting KB for Rumor Context Expansion (RQ3)*

To address RQ3, we expanded the rumor with the three approaches discussed in Section 3.6, and conducted an ablation study to assess the effect of combining those techniques on the performance of the *initial* and *hybrid* models, adopting *bio+lists* representation and *Arabic BERT* reranking models.

Table 3.8. Performance of initial retrieval with rumor expansion. Initial (*bio+lists*) model was used for retrieval. Symbol $\star$ indicates statistically significant difference compared to the initial retrieval with raw tweet baseline.

| **Query** | **P@1** | **P@5** | **NDCG@5** | **R@100** |
|---|---|---|---|---|
| *Raw Rumor* [**Initial**] | 0.373 | 0.167 | 0.238 | 0.425 |
| *+Wikipedia title* | **0.387** | **0.185** | **0.256** | 0.445 |
| *+DBpedia entities* | 0.313 | 0.140 | 0.201 | 0.407 |
| *+Wikipedia2vec entities* | 0.267 | 0.163 | 0.212 | 0.448 |
| *+Wikipedia title+DBpedia* | 0.347 | 0.163 | 0.233 | 0.445 |
| *+Wikipedia title+Wikipedia2vec* | 0.313 | 0.163 | 0.221 | 0.448 |
| *+DBpedia+Wikipedia2vec entities* | 0.320 | 0.167 | 0.233 | 0.477$^\star$ |
| *+all* | 0.333 | 0.175 | 0.244 | **0.488**$^\star$ |

Table 3.8 presents a comparison between *initial* retrieval with no expansion (i.e., using the *raw* tweet expressing the rumor as the query) and the same model when we

expand the *raw* tweet using various approaches. As expected, expanding the rumors with related entities improved the recall over the non-expansion baseline. In particular, expansion with DBpedia plus Wikipedia2vec entities and expansion with all approaches combined yielded a statistically significant improvement of 12.24% and 14.82%, respectively, in R@100 over the baseline. However, only expansion with Wikipedia titles outperformed the baseline in precision. In fact, expansion with Wikipedia titles improved the performance in all measures, but with no statistically significant difference.

Table 3.9. Performance of hybrid retrieval with rumor expansion. Hybrid (*bio+lists*) with Arabic BERT was used for retrieval. Symbol $\star$ indicates statistically significant difference compared to the hybrid retrieval with raw tweet baseline.

| **Query** | **P@1** | **P@5** | **NDCG@5** | **R@50** |
|---|---|---|---|---|
| *Raw Rumor* [**Hybrid**] | **0.413** | 0.213 | **0.271** | 0.396 |
| +*Wikipedia title* | 0.340 | 0.200 | 0.249 | 0.432$^\star$ |
| +*DBpedia entities* | 0.267 | 0.181 | 0.226 | 0.389 |
| +*Wikipedia2vec entities* | 0.340 | 0.211 | 0.253 | 0.422 |
| +*Wikipedia title*+*DBpedia* | 0.373 | 0.200 | 0.265 | 0.429 |
| +*Wikipedia title*+*Wikipedia2vec* | 0.293 | 0.197 | 0.229 | 0.423 |
| +*DBpedia*+*Wikipedia2vec* | 0.360 | **0.224** | 0.262 | 0.455$^\star$ |
| +*all* | 0.353 | 0.215 | 0.264 | **0.468**$^\star$ |

In Table 3.9, we present a similar comparison with *hybrid* retrieval. As a baseline for this experiment, we use the best hybrid model with no expansion in terms of P@5 (i.e., using *bio+lists* representation with Arabic BERT), as indicated in Table 3.6. We also report performance in R@50 to check if the expansion has a potential for further reranking. As the objective of the hybrid retrieval is precision at top ranks, we focus first on the precision performance. We first note that none of the expansion approaches outperform the baseline in P@1; this is somewhat expected as expansion methods typically hurt precision. However, both expansion with DBpedia plus Wikipedia2vec entities and expansion with all approaches combined have a slight improvement (that is not statistically significant) over the baseline in P@5 and comparable performance in NDCG@5. In terms of recall, we observe similar performance to the initial retrieval results, but now at lower ranks of 20 and 50; all expansion approaches (except expansion solely with DBpedia entities) outperform the baseline in both R@20 and R@50. In particular, expansion with DBpedia plus Wikipedia2vec entities and expansion with all approaches combined yield statistically significant improvements of 14.9% and 18.18%, respectively, in R@50 over the baseline. This shows expansion approaches are able to push good authorities higher but not to the very high ranks of the top 5, which indicates a clear potential for further reranking.

In Table 3.10, we present an example of a rumor along with a subset of its expansion entities and retrieved authorities. The results demonstrate that the hybrid model using the raw tweet as a query retrieved only one authority out of ten found by the annotators, as opposed to *eight* authorities retrieved when the rumor was expanded by a set of entities extracted from the KB and KG. Some of the *very relevant* expansion entities are: “Al-Ahly Sporting club Egypt”, “Wydad Athletic Club”, “Category: Egypt Cup seasons in 20st century”, and “Category: National association football team records

and statistics" (from KB), and "2017 CAF Champions League Final", "2017 CAF Champions League knockout stage", and "2019 CAF Champions League Final" (from KG).

To answer **RQ3**, we conclude that rumor expansion using the KB can significantly improve the recall of authorities but not precision. This makes it a good starting point to further improve the precision in future studies.

Table 3.10. Hybrid retrieval of authorities *with* and *without* rumor expansion with entities from KB and KG. Initial (bio+lists) model and Arabic BERT were adopted. Italic user names are *relevant* and *double underlined* ones are highly relevant.

**Rumor** Moroccan reports: Bakary Gassama is the referee for the return match between Al-Ahly and Wydad.

| | |
|---|---|
| **Retrieval Without Rumor Expansion** | |
| 2. @WACofficiel | The official account of Wydad Athletic Club |
| **Retrieval With Rumor Expansion** | |
| 1. @caf_online_AR | Confederation of African Football (CAF) |
| 3. @AlAhly | The official account of AlAhly club |
| 4. @CAFCLCC_ar | Africa's interclub competitions |
| 5. *@fifacom_ar* | The official FIFA account in Arabic |
| 10. *@CAF_Media* | CAF Communications Division |
| 12. @bibo | The Chairman of the Board of Directors of Al-Ahly Club |
| 27. *@AlAhlyTV* | Al Ahly TV official account |
| 56. @WACofficiel | The official account of Wydad Club |

### *3.7.4. Comparing with SOTA Topic Expert Finding Models (RQ4)*

Table 3.11. Our proposed models vs. SOTA topic expert finding models. A star indicates statistically significant difference compared to Cognos model.

| **Model** | **P@1** | **P@5** | **NDCG@5** | **R@100** |
|---|---|---|---|---|
| *Cognos* | 0.173 | 0.085 | 0.114 | 0.336 |
| *FAME* | 0.040 | 0.029 | 0.034 | 0.158 |
| *CombMNZ* | 0.053 | 0.031 | 0.036 | 0.056 |
| Lexical (*lists*) | 0.193 | 0.103 | 0.141 | 0.329 |
| Lexical (*bio* + *lists*) | 0.180 | 0.097 | 0.119 | 0.318 |
| Initial (*lists*) | 0.347$^\star$ | 0.151$^\star$ | 0.221$^\star$ | **0.428**$^\star$ |
| Initial (*bio* + *lists*) | 0.373$^\star$ | 0.167$^\star$ | 0.238$^\star$ | 0.425$^\star$ |
| Hybrid (*bio* + *lists*, Arabic BERT) | **0.413**$^\star$ | **0.213**$^\star$ | **0.271**$^\star$ | 0.425$^\star$ |

Finally, we compare the performance of our proposed models with SOTA models of topic expert finding in social media, as identified in Section 3.6. The results presented

in Table 3.11 yield multiple insights. First, among the SOTA models, Cognos [139] is clearly the best performing in all measures over AuFIN. In fact, the other two (FAME and CombMNZ) exhibit very poor performance (explained below). Second, the performance of Cognos is comparable to our *lexical* models using *bio* and *lists* text representation. Finally, our initial and hybrid models significantly outperform Cognos, the best SOTA model, in all evaluation measures.

We also investigated the poor performance of both FAME and CombMNZ over AuFIN. As for FAME [143], the model depends on the construction of a user endorsement graph, where a user $u_i$ endorses another user $u_j$ if $u_i$ created a list and added $u_j$ as a member of that list. In the original implementation of FAME, the endorsement graph was constructed *iteratively* using a seed of experts for *only three* pre-specified topics. This resulted in a graph of 139,798 users that is *closed* in terms of endorsement. In our implementation, we created our endorsement graph using our *full/existing* user collection of 291,653 users (users listed in Twitter lists). Due to the large scale of the collection, this resulted in missing some endorsements (both nodes and edges in the graph) due to missing some list owners from our collection. In fact, our collection captures, on average, 30% of each user's real endorsements on Twitter. While this indeed limits the effectiveness of the model, Lahoti, De Francisci Morales, and Gionis [143] explicitly mentioned that "*our dataset is crawled from a specific set of seeds. The effect of this choice on the ranking algorithm needs further investigation.*", acknowledging the limitation.

As for CombMNZ [200], the model retrieves tweets that are potentially relevant to the rumor from all user timelines in our collection. By manually checking those retrieved tweets, we found that many are duplicates or near duplicates of the rumor, which means that the model is actually retrieving users who propagated the same rumor, not authorities, yielding poor performance for our task.

To answer **RQ4**, we found that SOTA models of expert finding in social media performed poorly for our authority finding task over AuFIN, for different reasons we explained. Moreover, our hybrid model, in particular, significantly outperforms those SOTA models in both precision and recall measures.

## 3.8. Discussion

In this section, we discuss our evaluation results in terms of failure cases (Section 3.8.1) and limitations (Section 3.8.2).

### *3.8.1. Failure Analysis*

We conducted a detailed error analysis on all rumors for which our *hybrid* model (*bio + lists and Arabic BERT*) could not retrieve any authority at depth 100. This constitutes 22 rumors expressed in tweets (15% of our set). We study the two types of errors, i.e., *false positives* and *false negatives*, and categorize the reasons behind these errors based on a thorough examination of the failed examples. We also discuss why the timeline user representation, in particular, fails to achieve good performance.

#### *3.8.1.1. False Positives*

Table 3.12 presents example rumors along with their top retrieved *false positive* user. We list below the main reasons behind those errors and highlight them in the table:

- **Non-focused tweets:** Some users propagating rumors tend to augment the rumor with their own opinion, analysis, or sarcasm, which adds noisy or irrelevant terms to the actual claim, e.g., $T_2$. This emphasizes the need for accurate claim extraction to filter out irrelevant terms.

- **Misleading hashtags or terms:** In our work, we are not excluding hashtags as they are, sometimes, part of the rumor text; however, some hashtags are actually misleading, e.g., $T_4$ in Table 3.13 which contains "#AlShula_electronic_newspaper" hashtag that is irrelevant to the rumor. This misled the model to retrieve Twitter accounts offering electronic learning, training, and education. Similarly, some rumors could have some terms that may lead to retrieving irrelevant users. The tweet $T_2$ in Table 3.12 mentions "glass" multiple times, which led to retrieving the Twitter account of a thermal insulation of glass company.

- **Sense ambiguity:** Some rumors mention a term that can have different meanings (senses) with the same spelling, e.g., $T_1$, which mentions the term "jamal" that may denote a name of a person or mean beauty in Arabic.[35]

- **Non-Arabic terms:** Some rumors mention entities in non-Arabic language, e.g., $T_2$ that mentions "Schott" and "Mayence." In our work, we exclude any non-Arabic terms, which sometimes constitute key terms in the claim.

[35]Jamal and beauty have different spellings in the English translation, but, in Arabic, they have exactly the same spelling.

Table 3.12. Sample *false positive* cases of (translated) rumors and user representations. The lexical overlap, the misspelled entities, and non-Arabic terms are underlined, double underlined, and triple underlined. italicized are the entities mentioned in the rumor.

| **Tweet (Failure Reason(s))** | **Retrieved non-Authority & Textual Representation** |
|---|---|
| $T_1$ **(Sense ambiguity, Misspelling)**<br>*Jamal Belamri* donates 500 oxygen relaxers to hospitals, which will be sent on Friday. Jamal may god gives you great health, you are our pride | **@alvitaminblog**<br>Vitamin. Towards a better healthy life. Health, healthy, medical sciences, self help and health, health and fitness, vitamins, health and beauty, health and medicine, health and medical information, education health, health and beauty, health, health, healthy, medical, life ... |
| $T_2$ **(Non-focused, Misleading terms, Non-Arabic terms)**<br>According to *German police* investigations: *Pfizer* ordered from the German company *Schott*, located in *Mayence*, which is specialized in making glass equipment for laboratories, two orders, each one for 800 million glass bottles, with special glass that can withstand a temperature of minus 100 degrees Celsius. This was on November 2, 2019 Before the start of Cosfit | **@xeoex_sa**<br><br>Xeoex. Al-Bazai Group is a company specialized in thermal insulation of glass and paint protection. The agent in the Kingdom and the Gulf countries is Al-Bazai Group. There are no branches or distributors outside Riyadh. Other, projects. |

*3.8.1.2. False Negatives*

Table 3.13 presents example rumors along with a *false negative* authority, which our model failed to retrieve within the top 100. We discuss below the main factors behind those errors and highlight them in the table:

- **Lack of context:** Some rumors may need additional context in order to retrieve the corresponding authorities. For example, $T_3$ is a rumor about "Ataq Hospital" located in Shabwa, a province of Yemen. Since it did not mention "Yemen", the model failed to retrieve any *Yemeni* health authorities.
- **No (or low) lexical overlap:** Our model retrieves the set of initial users by lexical matching; however, we found that some rumors do not have any lexical overlap

with the authority representation, e.g., $T_3$. Moreover, some may have very low lexical overlap, especially if the user representation is short, e.g., $T_4$.

- **Misspelling:** Some mentioned terms or entities are misspelled, e.g., $T_3$ in Table 3.13 (and $T_1$ in Table 3.12).
- **Low list count:** Users listed in few or no Twitter lists are less likely to be retrieved, e.g., $T_4$. This can be attributed to two reasons. First, the model initially relies on lexical matching with list-based user representation. Second, list count is a major factor of the *initial* scores, which gives "priority" to users with a high number of lists.

Table 3.13. Sample *false negative* cases of (translated) rumors and user representations. The lexical overlap, and the misspelled entities are highlighted in underlined and double underlined. italicized are the entities mentioned in the rumor.

| **Tweet (Failure Reason(s)** | **Non-retrieved Authority (list count) & Textual Representation** |
|---|---|
| $T_3$ **(No lexical overlap, Low list count, Misspelling, Lack of context)**<br>If this news is true that they are transporting the blocks in this way in the new *Ataq Hospital*. There is a real development in the possibility of transferring Cororna victims helicopters for treatment abroad, at the expense of the local authority in *Shabwah* | **@YEMEN_MOH (15)**<br><br>Ministry of Public Health and Population-Republic of Yemen. The official account of the Yemeni Ministry of Public Health and Population. Saudi Yemen group, Yemen, Yemen news, Yemen Corona Response, Coronavirus, Corona, Corona, Corona, Corona Yemen news, Yemen, Government agencies, Media. |
| $T_4$ **(Low list count, Low lexical overlap)**<br>Reports: Cancellation of the *Moroccan League* due to the refusal of the competent authorities to complete the season #AlShula_electronic_newspaper | **@ElhadaouiM (6)**<br>Mustapha El Hadaoui, President of the Moroccan Association of Footballers AMF and CAF & FIFA Instructor,Moroccan footballers a list consisting of Moroccan professionals, past and present, former Players now they do other things or nothing,Morocccan football trade unionists it is the defense of the interests of footballers in the world, sport. |

#### *3.8.1.3. Failure of Timeline Representation*

Due to the poor performance when incorporating the *timeline* user representation, we were curious to investigate the primary cause behind it. We manually examined the top retrieved users for some rumors using the timeline representation. While many terms of the rumor could appear within the timeline, they appear individually *scattered* over the timeline in several and different contexts, yielding a high lexical score that is not really indicative of relevance. Furthermore, in some cases, the exact rumor might appear in the timeline as is, indicating the user is just "spreading" the rumor rather than being an authority for it. Even when *lists* are added to the timeline, the same matching of scattered terms occurs, yielding almost no difference in the retrieval results. Overall, the timeline reflects the attitude and activities of users regarding different topics, thus it is very challenging to get a clear authority signal out of that.

### *3.8.2. Limitations of Our Study*

The methodological limitations of our study are related to both AuFIN and the proposed model.

- **Test Collection:**
  - In terms of associating authorities with rumors, this relies, in some cases, on the annotators' background and subjectivity, which may lead to an incomplete set of authorities.
  - AuFIN is relatively small in size as the annotations were time-consuming. The annotators needed to search for the names of authorities related to specific domains in specific countries, then look for their Twitter accounts if they exist, which requires a considerable amount of time. Annotators reported an average of 45 minutes to annotate a single rumor, 2 to 3 hours per 10 rumors to check agreements among annotators, and even further time to resolve final disagreements by a third one.
- **Proposed Model:**
  - Depending on Twitter lists as the main source of the textual representation of users for authority finding proved, in our work, to be *insufficient*, especially since our model adopts lexical retrieval to get an initial set of candidates. This is despite the fact that it was shown to be effective for topic expert finding [139], [143]. The major difference is that the queries in our case are much longer, with several terms expressing the rumor that might not appear in the list names and descriptions, compared to a few terms expressing the topic for expert finding.
  - Our models use the tweets mostly as is, with little pre-processing. That led (as illustrated in Section 3.8.1) to matching irrelevant terms, thus retrieving non authorities.
  - The model gives higher priority to users with a higher number of Twitter lists and followers. While this improved the results significantly as opposed to lexical retrieval solely, some authorities in reality are listed in few lists and have a low ratio of followers to followees.

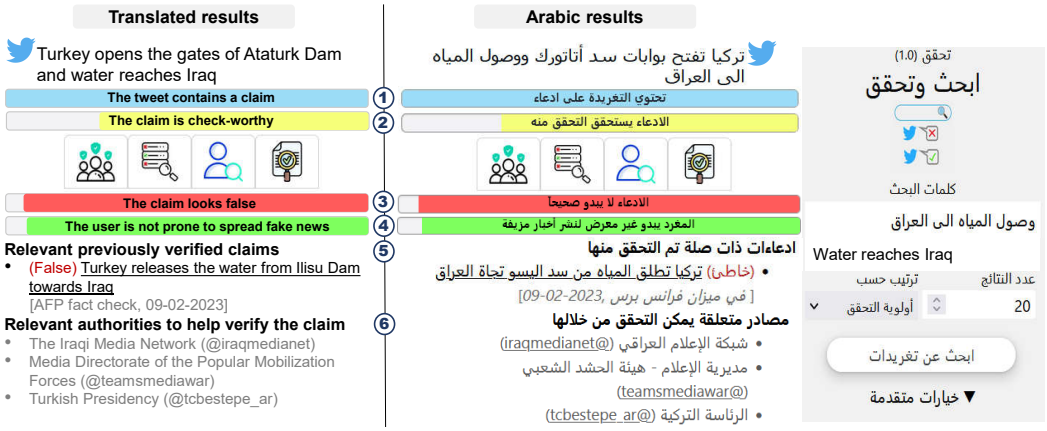



Figure 3.5. Example of a retrieved tweet in Tahaqqaq with output of each component: (1) Claim Identification, (2) Check-worthy Claim Detection, (3) Claim Verification, (4) User Credibility, (5) VClaim Retrieval, and (6) *Authority Finding*.

– All the query expansion approaches adopted in our work rely on concatenating relevant entities from the KG to the rumor text. We believe we need to further explore techniques for selecting the "best" entities to augment the rumor in order to retrieve the relevant authorities rather than using all of them for expansion.

The above limitations motivate the need to (1) having multiple annotators, preferably from different backgrounds, to ensure better coverage of the set of authorities, (2) exploring other techniques and other Twitter features to represent users, as relying on lexical matching and lists for authority finding is clearly sub-optimal, (3) adopting an effective prepossessing pipeline, including filtering irrelevant hashtags and terms, accurately extracting the key terms of the rumors for a more effective rumor representation, correcting misspellings, and translating non-Arabic terms, and (4) exploring methods for selecting relevant entities for rumor expansion to improve the performance of the model.

## 3.9. Model Deployment

Our proposed authority finding model was deployed as part of Tahaqqaq [33], a live demo for Arabic claim verification with multiple components, to retrieve authorities for a given tweet retrieved in real-time or any free-text claim. Figure 3.5 presents an example of a retrieved tweet along with the output of Tahaqqaq components including the authority finding model indicated as number 6. For clarity, the figure also shows the English translation of the search query and all prediction results.

## 3.10. Conclusion

In this chapter, we introduced the authority finding task in Twitter that can help both fact-checkers and automated systems in finding the authoritative Twitter accounts for specific rumors, hence helping in the verification process. We constructed and released the first test collection for Arabic authority finding in Twitter to enable fur-

ther studies on the task and shared our language-independent annotation guidelines to encourage the construction of similar collections in other languages. We proposed a hybrid model that employs pre-trained language models and combines lexical, semantic, and network signals to find authorities. Moreover, we explored the effect of expanding the context of the rumors with entities retrieved from the Knowledge Base. Our experimental results suggest that Twitter lists and network features, adopted previously for topic expert finding models, play a crucial role in authority finding; however, they are insufficient. Results also showed that semantic ranking of candidate authorities improved the recall of authorities, but degraded the precision. However, combining all signals improved the precision significantly. We further showed that the expansion of rumors by relevant entities significantly improved the recall of authorities. Finally, out of our thorough failure analysis, we recommend further work on the pre-processing pipeline, exploring other sources for representing users, and exploiting other features to differentiate experts from authorities.

In this chapter we addressed finding authorities who can help verify a specific rumor. Incorporating those retrieved authorities for rumor verification will be addressed in the subsequent chapters. In Chapter 4, we target detecting the stance of the retrieved authority timeline tweets toward a specific rumor. We address evidence retrieval from retrieved authority timelines in Chapter 5. Finally, we show how evidence retrieved from authority timelines can be used for rumor verification in Chapter 6.

# CHAPTER 4: DETECTING STANCE OF AUTHORITIES TOWARDS RUMORS

In chapter 3, we addressed finding authorities who can help verify a specific rumor. In this chapter, we introduce one way of incorporating those retrieved authorities for rumor verification. Specifically, we propose detecting the stance of their timeline tweets as a signal for rumor verification.

A large body of existing studies in the broader literature have examined exploiting the stance of conversational threads [4], [13] or news articles [202], [203] towards claims as a signal for verification. However, to our knowledge, no previous research has investigated exploiting the stance of *trusted authorities* for rumor verification in social media. Therefore, we believe that detecting stance of relevant authorities towards rumors can be a great asset to augment the sources of evidence utilized by existing rumor verification systems. It can also serve as a valuable tool for fact-checkers to automate their process of verifying rumors from authorities.

This chapter starts with defining *detecting stance of authorities towards rumors* task in Section 4.1. We give an overview about our work in Section 4.2. In Section 4.3, we present our dataset construction approach. Our experimental approach is presented in Section 4.4. We discuss the experimental setup in Section 4.5 and thoroughly analyze the results and answer the research questions in Section 4.6. We conduct a failure analysis to gain insights for future directions and discuss the limitations of our study in Section 4.7. Finally, we conclude and suggest some future directions in Section 4.8.

## 4.1. Problem Definition

We propose the task of *detecting stance of authorities towards rumors* defined as follows: Given a rumor expressed in a tweet and a tweet posted by an authority of that rumor, detect whether the tweet *supports* (agrees with) the rumor, *denies* (disagrees with) it, or *not* (other).

## 4.2. Overview of Our Work

Figure 4.1 shows an example of a rumor about an establishment of a new railway to connect the Sultanate of Oman and the United Arab of Emirates (UAE). We assume that the authorities for this rumor are retrieved by an "authority finding" model (here some of the highly relevant authorities are the ministry of transport in Oman, the Omani government communication center, and both Oman's and UAE's rails projects). The figure shows an example tweet from each of the timelines of the authorities that actually supports the rumor.[1]

In this chapter, we introduce the task of detecting the stance of authorities towards rumors in Twitter. Due to the lack of datasets for the task, we construct and release the first Authority STance towards Rumors (AuSTR) dataset (Section 4.3). We exploit both fact-checking articles and authority Twitter accounts to *manually* collect *debunking*, *supporting*, and *other* (rumor tweet, authority tweet) pairs. Additionally, we propose a semi-automated approach utilizing the Twitter search API to further expand our *debunking* pairs.

Due to the limited size of our dataset, we investigate the usefulness of existing datasets of stance towards Arabic claims (Section 4.6.1 and Section 4.6.2). Adopting a BERT-based stance model, we perform extensive experiments using 5 variant Arabic

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

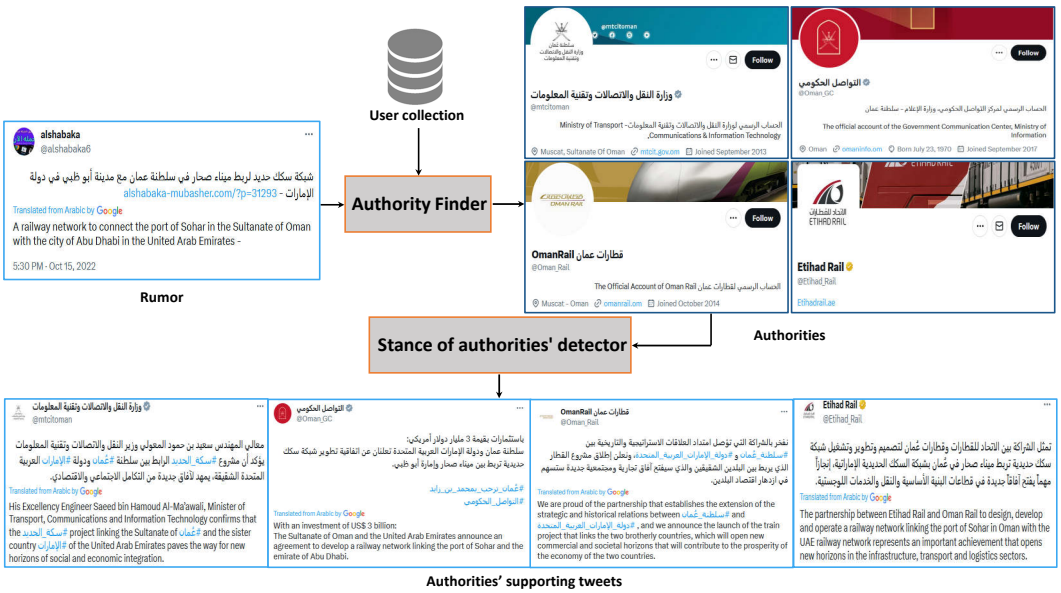



Figure 4.1. An example of a rumor along with its corresponding authorities and a set of *supporting* tweets detected from the authorities timelines (The example is from our constructed AuSTR dataset).

stance datasets, where the target is a claim but the context is either an article, article headline, or a tweet, to investigate if the stance model trained with each of them is able to generalize to our task. We then explore the effect of augmenting our in-domain data with each of the Arabic stance datasets on the performance of the model (Section 4.6.3). To mitigate the class-imbalance issue, we explore variant loss functions replacing the cross-entropy loss (Section 4.6.4). Finally, we conduct a thorough error analysis to gain insights for the future improvements (Section 4.7.1). The contributions of this chapter are as follows:

1. We introduce and define the task of detecting the stance of authorities towards rumors that are propagating in Twitter.

2. We release the first Authority STance towards Rumors (AuSTR) dataset for that specific task[5] targeting the *Arabic* language.

3. We explore the adequacy of existing Arabic datasets of stance towards claims for our task, and the effect of augmenting our in-domain data with those datasets on the performance of the model.

4. We investigate the performance of the models when adopting variant loss functions to alleviate the class-imbalance issue, and we perform a thorough failure analysis to gain insights for future work on the task.

## 4.3. Constructing AuSTR Dataset

To address the lack of datasets of authority stance towards rumors, in this work, we introduce the *first* Authority STance towards Rumors (denoted as *AuSTR*) dataset. Our focus is on Arabic, as it is one of the most popular languages in Twitter [150],

yet it is under-explored for rumor verification. Our dataset consists of 811 pairs of rumors (expressed in tweets) and authority tweets related to 292 unique rumors. Tweets of authorities are labeled as either *disagree*, *agree*, or *other*, as defined earlier. To construct AuSTR, we collected the *debunking* pairs manually (details in Section 4.3.1) by exploiting fact-checking articles and adopting a semi-automated approach. *Supporting* pairs were collected by manually exploring authority accounts and the Twitter search interface, in addition to utilizing the fact-checking articles (details in Section 4.3.2). Finally, to collect our *other* pairs we manually examined the timelines of the authorities of our *debunking* and *supporting* pairs to select tweets that are neither agreeing nor disagreeing with the rumor, in addition to exploiting fact-checking articles (details in Section 4.3.3).

### *4.3.1. Collecting Debunking Pairs*

Figure 4.2 depicts an overview of our approach to construct the *debunking* pairs of AuSTR. We leveraged both the fact-checking articles and a semi-automated approach which we propose in this work.

#### *4.3.1.1. Exploiting Fact-Checking Articles*

Fact-checkers who attempt to verify rumors usually provide, in their fact-checking articles, some examples of social media posts (e.g., tweets) propagating the specific rumors, along with other posts from trusted authorities that constitute evidence to support their verification decisions. For AuSTR, we exploit both examples of tweets: stating rumors and showing evidence from authorities as provided by those fact-checkers. Specifically, we used AraFacts [204], a large dataset of Arabic rumors collected from 5 fact-checking websites. From those rumors, we selected only the ones that are expressed in tweets and for which the fact-checkers provided evidence in tweets as well.[2]

For *false* rumors, we selected a single tweet example of the rumor and all provided evidence tweets for it, which are then labeled as having *disagree* stances. Adopting this approach, we ended up with 118 *debunking* pairs.

#### *4.3.1.2. Exploiting Twitter Search*

Additionally, we adopted a semi-automated approach to collect more debunking pairs using Twitter search. First, we used the Twitter Academic API[3] to collect *potentially-debunking* tweets, i.e., tweets with denying keywords and phrases such as "*fake news*," "*fabricated*, *rumors*," and "*denied the news*." Specifically, we used 21 keywords/phrases[4] to search Twitter to retrieve Arabic tweets from the period of July 1, 2022 to December 31, 2022. To narrow down our search and reduce the noisy tweets, we excluded retweets and the tweets of non-verified accounts. Given that fact-checkers usually use most of these keywords to debunk rumors, we also excluded tweets from verified Arabic fact-checking Twitter accounts.

By adopting this approach, we were able to collect either debunking tweets from authorities themselves, or just *pointer* tweets from journalists or news agencies. For both types, we retrieved the rumor tweets by searching Twitter user interface using the main

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

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

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

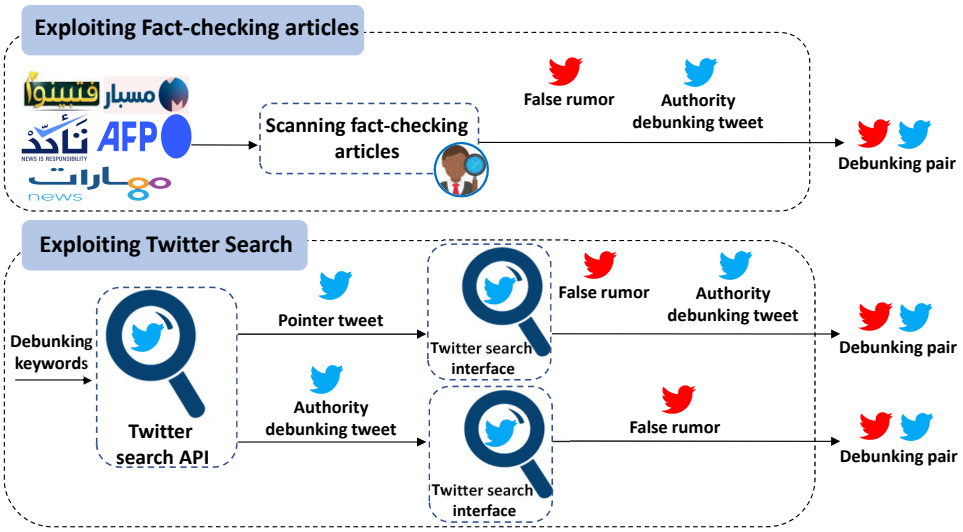



Figure 4.2. Our approach for collecting AuSTR *debunking* pairs.

keywords in the debunked rumor by the authorities. For the later type, we manually examined the timelines of authorities to get the debunking tweets.

Table 4.1 presents examples of *debunking* tweets from authorities along with the search keywords used to retrieve them (translated to English). An example of automatically-retrieved pointer tweet and the manually-collected *disagree* pair is presented in Table 4.2.

Table 4.1. Examples of *debunking* authority tweets (their English translation) collected using the semi-automated approach along with the search keywords.

| Search keywords | Example of a collected tweet |
|---|---|
| Incorrect | @**AymanNour:** Statement from #Ghad El Thawra: One of the sites published incorrect news about the party's decision to call for the 11/11 movement ... |
| Fake news | @**LebISF:** Denying a fake news published by a Lebanese newspaper about the arrest of Major General Othman's brother |
| Untrue | @**IraqiSpoxMOD:** ... news about (the disappearance of an American citizen in central or southern Iraq, under mysterious circumstances, who works as a journalist). We confirm that this news is untrue ... |
| Fabricated | @**AlAhlyTV:** ...Al-Ahly's objection speech about Zamalek club uniforms in the super is fabricated... |
| Rumors | @**DGSGLB**: #Statement: rumors are circulating that the General Directorate of General Security arrested Sally Hafez, who broke into a bank in Beirut... |

Table 4.2. An example of an automatically collected *pointer debunking* tweet along with its manually collected *debunking* pair (their English translation).

| Tweet type | Tweet text |
|---|---|
| **Pointer** | **@naharkw**: The Qatari Embassy in Tunisia: Incorrect.. A Qatari was killed in the ancient city of Bizerte. [11-08-2022] |
| **Authority** | **@QatarEmb_Tunis:** The Embassy of the State of Qatar in the Republic of Tunisia denies what was reported by the media that the victim in the Bizerte incident holds Qatari nationality, and expresses its condolences to the victim's family and relatives. [11-08-2022] |
| **Rumor** | **@USER:** The killing of a Qatari in Tunisia shakes the ancient city of Bizerte #Tunisia [12-08-2022] |

Table 4.3. An example of manually collected *supporting* authority tweet and a relevant rumor tweet expressing the same claim (their English translation).

| Tweet type | Tweet text |
|---|---|
| **Authority** | **@Moi_kuw:** A resident who tried to commit suicide by stabbing himself inside a mosque was first aided, and the person was kept and the necessary legal measures are being taken in the incident. [04-12-2022] |
| **Rumor** | **@USER:** Circulating #suicide_attempt: He attempted suicide inside Al-Ghanim Mosque in Cordoba, and the reasons are still unknown.[04-12-2022] |

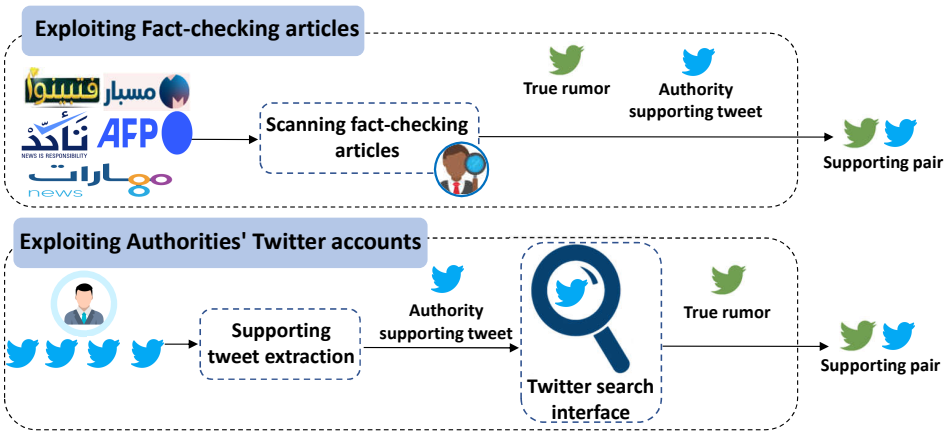



Figure 4.3. Collecting AuSTR *supporting* pairs approach.

### *4.3.2. Collecting Supporting Pairs*

To collect *supporting* pairs, we adopted two approaches as presented in Figure 4.3. Given that fact-checkers focus more on *false* rumors than *true* ones, exploiting fact-checking articles was not sufficient to collect *supporting* tweets, as adopting this approach, we were able to collect only 4 *agree* pairs as opposed to 118 *disagree* pairs. Thus, we manually collected a set of *governmental Arabic Twitter accounts* representing authorities related to health and politics, such as ministries and ministers, embassy accounts, and Arabic Sports organizations accounts (e.g., football associations and clubs). Starting from 172 authority accounts from multiple Arabic countries,[5] we manually checked the timelines of those authorities from the period of July 1, 2022 to December 31, 2022. We selected *check-worthy* tweets, i.e, tweets containing verifiable claims that we think will be of general interest [47], and consider them as authority *supporting* tweets. We then used the main keywords in each claim to search Twitter through the user interface and selected a tweet propagating the same claim while avoiding near-duplicates. We ended up with 148 *agree* pairs in total. Table 4.3 shows an example of a *supporting* authority tweet along with a relevant rumor.

### *4.3.3. Collecting Other Pairs*

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

In reality, most of the tweets in authority timelines are neither supporting nor denying a given rumor. To get closer to that real scenario,

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

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

### *4.3.4. Data Quality*

We present our dataset statistics in Table 4.4. Our data was annotated by one of the authors, a PhD candidate and native Arabic speaker working on rumor verification in Twitter. To measure the quality of our data, we randomly picked 10% of the pairs and asked a *second* annotator, a PhD holder and native Arabic speaker, to label them. The computed Cohen's Kappa for inter-annotator agreement [158] was found to be 0.86, which indicates "almost perfect" agreement [159].

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

Table 4.4. AuSTR statistics.

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

## 4.4. Experimental Design

Due to the limited size of AuSTR, one of the main objectives of this work is to study the adequacy of using *existing* datasets of stance towards claims in training models for our task. Specifically, the goal is to first study whether models trained with existing stance datasets perform well on detecting the stance of authorities in particular, then investigate whether augmenting them with AuSTR improve the performance of those models. Moreover, since a major challenge of stance classification is the class-imbalance problem in the data [205], we also aim to explore whether incorporating different loss functions can mitigate that issue to further improve the performance of the models.

Accordingly, we aim to answer the following research questions:

- **RQ5:** To what extent will stance models trained with existing stance datasets be able to generalize to the task of detecting the stance of *authorities*? (Section 4.6.1)
- **RQ6:** What is the effect of combining all existing stance datasets for training? (Section 4.6.2)
- **RQ7:** Will training a stance model with AuSTR solely be sufficient? will augmenting AuSTR with existing stance datasets for training improve the performance? (Section 4.6.3)
- **RQ8:** Will adopting different loss functions mitigate the class-imbalance problem thus improve the performance? (Section 4.6.4)

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

- **Cross-domain experiments** denote the case where existing datasets of stance towards claims are exploited for training. Each of the stance datasets is first used solely for training our models, then all datasets were aggregated and used for training. We refer to the datasets of stance towards claims as *cross-domain* datasets in the rest of the paper.
- **In-domain experiments** denote the case where AuSTR is used solely for training. We refer to AuSTR as *in-domain* dataset.

- **In-domain augmented experiments** denote the case where AuSTR is augmented with existing datasets of stance towards claims. In those experiments, we study the effect of augmenting AuSTR with each of the cross-domain datasets separately, in addition to augmenting it with all of them.
- **Class-Imbalance experiments** denote the case where we adopt different loss functions, that showed promising results earlier in the literature, to alleviate the class-imbalance problem.

## 4.5. Experimental Setup

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

### *4.5.1. Datasets*

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

- **ArCOV19-Rumors [7]** consists of 9,413 **tweets** relevant to 138 COVID-19 Arabic **rumors** collected from 2 Arabic fact-checking websites. We considered the tweets *expressing the rumor* as supporting (agree), the ones that are *negating the rumor* as denying (disgree), and the ones discussing the rumor but neither expressing nor negating it as *other*.
- **STANCEOSAURUS [206]** consists of 4,009 **(rumor, tweet)** pairs. The data covers 22 Arabic rumors collected from 3 Arabic fact-checking websites along with tweets, collected by the authors, that are relevant to the rumors. The relevant tweets were annotated by their stance towards the rumor as either *supporting* (agree), *refuting* (disgree), *discussing*, *querying*, or *irrelevant*. In our work, we considered the last three labels as *other*.
- **ANS [207]** consists of 3,786 **(claim, manipulated claim)** pairs, where claims were extracted from news article headlines from trusted sources, then annotators were asked to generate *true* and *false* sentences towards them by adopting paraphrasing and contradiction respectively. The sentences are annotated as either *agree*, *disagree*, or *other*.
- **ArabicFC [208]** consists of 3,042 **(claim, article)** pairs, where claims are extracted from a single fact-checking website verifying political claims about the war in Syria, and articles collected by searching Google using the claim. The articles are annotated as either *agree*, *disagree*, *discuss*, or *unrelated* to the claim. In our work, we considered the last two labels as *other*.
- **AraStance [203]** consists of 4,063 **(claim, article)** pairs, where claims are extracted from 3 Arabic fact-checking websites covering multiple domains and Arab countries. The articles were collected and annotated similar to ArabicFC.

Figure 4.4 presents the per-class statistics for each dataset (including AuSTR), and Table 4.5 shows an example of a debunking text from each of them.

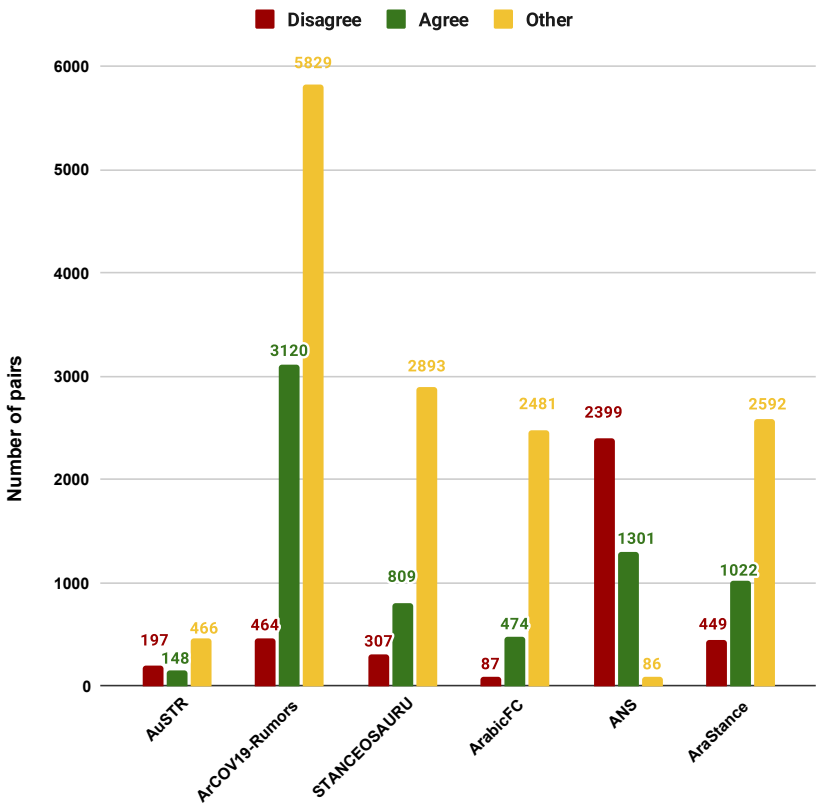



Figure 4.4. Per-class statistics of cross-domain datasets adopted in our work, as well as AuSTR for comparison.

Table 4.5. *Debunking* examples (their English translations) from the cross-domain datasets.

| **Dataset** | **Tweet text** |
|---|---|
| **ArCOV19-Rumors** | **@USER:** There is no truth to what is being circulated about Juventus player Paulo Dybala being infected with the Corona virus, and the source of the rumor is a Venezuelan channel. [13-03-2020] |
| **STANCEOSAURUS** | **@USER:** I was crying over the death of Kadim Al Sahir and it turned out that who died is his brother. [13-01-2022] |
| **ANS** | The Moroccan judiciary issued a 20-year prison sentence for Zefzafi. |
| **ArabicFC** | Hayat Tahrir al-Sham denies that its commander al-Julani was injured in a Russian strike, Al-Jazeera Mubasher, Wednesday, October 4, 2017... |
| **AraStance** | The circulating video entitled “a mobile phone explosion in a person’s pocket in a Dubai mall” is not true. Rather, it happened a few days ago in the city of Agadir in Morocco... |

#### *4.5.2. Data Splits*

Given that AuSTR constitutes only 811 pairs, we adopt cross-validation for evaluating our models. We randomly split it into 5 folds while assigning all pairs that are relevant to the same rumor to the same fold to avoid label leakage across folds. For all of our models, whether AuSTR is exploited for training or not, we both *tune* and *test* only on folds from AuSTR; a single AuSTR fold (dev fold) is used for tuning the models and another (test fold) was used for testing. If AuSTR is used for training, the remaining 3 folds (training folds) are used for that purpose. When the cross-domain datasets are used for training, they are fully used for that purpose (and none of them are used for tuning nor testing). For each experiment, we train 5 models to test on the 5 different folds of AuSTR, and finally report the average performance of the five models.

#### *4.5.3. Stance Models*

To train our stance models, we fine-tuned BERT [95], following recent studies that adopted transformer-based models for stance detection [203], [206], [209], [210] to classify whether the evidence *agrees* with the claim, *disagrees* with it, or *other*. We feed BERT the claim text as sentence *A* and the evidence as sentence *B* (truncated if needed) separated by the [SEP] token. Finally, we use the representation of the [CLS] token as input to a single classification layer with three output nodes, added on top of BERT architecture to compute the probability for each class of stance.

Various Arabic BERT-based models were released recently [167]–[169], [187], [188]; we opted to choose ARBERT [168] as it was shown to achieve better performance on most of the stance datasets adopted in our work [203]. All models were trained with a maximum of 25 epochs where 5 was set as an early stopping threshold. We tuned our models by adopting three variant learning rates (1e-5, 2e-5, 3e-5). The sequence length and batch size were set to 512 and 16 respectively.

#### *4.5.4. Preprocessing*

We processed all the textual content by removing non-Arabic text, special characters, URLs, diacritics, and emojis from the tweets. For STANCEOSAURUS, we extended the tweets with their context as suggested by the authors [206] who showed that extending the tweets with parent tweet text and/or embedded articles titles can improve the performance of the stance models.[6]

#### *4.5.5. Loss Functions*

We adopted the Cross Entropy ($CE$) loss in all our experiments. However, due the imbalanced class distribution, we also experimented with the Weighted Cross Entropy ($WCE$) loss, and Class-Balanced Focal ($CBF$) loss [211] adopted by Baheti, Sap, Ritter, *et al.* [212] and Zheng, Baheti, Naous, *et al.* [206] to mitigate the issue for stance detection. For $CBF$, we set the hyperparameters $\beta$ and $\gamma$ to 0.9999 and 1.0 respectively as suggested by Baheti, Sap, Ritter, *et al.* [212].

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

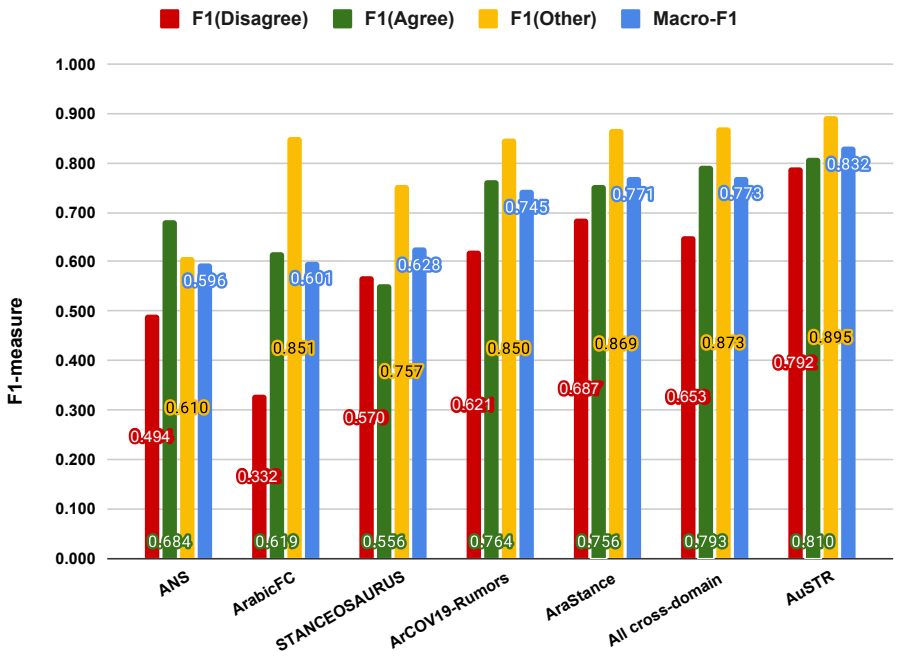



Figure 4.5. The performance of models trained using *cross-domain* vs. *in-domain* datasets.

### *4.5.6. Evaluation Measures*

To evaluate our models, we report the average of macro-$F_1$ scores across the 5 folds of AuSTR, in addition to average per-class $F_1$. Macro-$F_1$ is recommended to evaluate stance models [213] due to the class-imbalance nature of stance datasets

## 4.6. Experimental Evaluation

In this section, we present and discuss the results of our experiments that address the research questions introduced in Section 4.4.

### *4.6.1. Leveraging Cross-domain Datasets for Training (RQ5)*

To address **RQ5**, we used the five cross-domain datasets listed earlier for training. For each of them, we train on the full cross-domain dataset, then fine-tune 5 stance models; each is tuned on one fold from AuSTR and tested on another fold. We report the average performance on testing on the 5 folds of AuSTR in Figure 4.5.

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

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

exhibits the best performance.

As for the inferior category, we speculate the rationale behind their performance. We note that ArabicFC is severely imbalanced, where the *disagree* class represents only 2.86% of the data, yielding a very poor performance on that class. Moreover, it covers claims related to only one topic , which is the Syrian war, making it hard to generalize. A similar conclusion was found by previous studies that used ArabicFC [203], [208].

As for ANS, evidence was manually/artificially crafted, which is not as realistic as tweets from authorities. As for STANCEOSAURUS, it covers tweets relevant to only 22 claims.

As for the superior category, we observe that AraStance and ArCOV19-Rumors achieved the highest $F_1$ on the *disagree* class compared to the other cross-domain datasets. ArCOV19-Rumors covers 138 COVID-19 claims in several topical categories. AraStance covers 910 claims, which are extracted from three fact-checking websites, covering multiple domains and Arab countries, similar to AuSTR, and the evidence is represented in articles written by journalists, not manually crafted. To further investigate their performance, we manually examined 20% of AraStance and ArCOV19-Rumors *disagreeing* training pairs. We found that about 68% and 59% of the examined examples of AraStance and ArCOV19-Rumors respectively share common debunking keywords, such as "*rumors*," "*not true*," "*denied*," and "*fake*;" similar keywords appear in some *disagreeing* tweets of AuSTR.

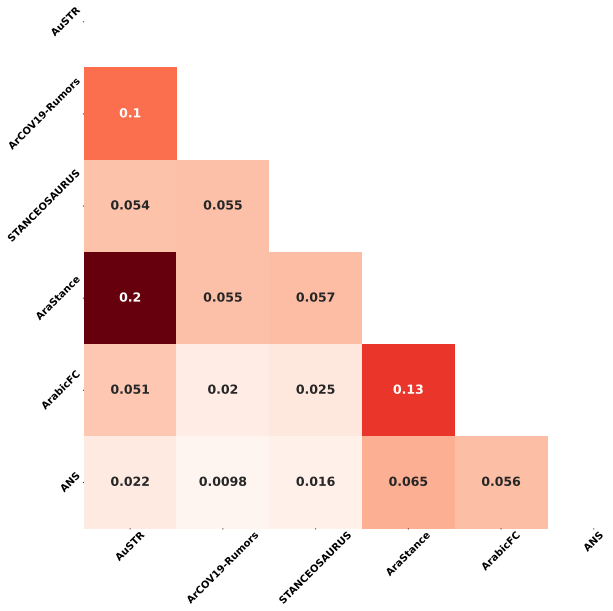
To further investigate the relation between the datasets and the performance of the corresponding models, we analyzed the lexical similarity between the datasets. We first constructed a 2-gram vector representation for each dataset (including AuSTR) using the preprossessed context[7](excluding the claims), then we performed a pairwise cosine similarity between the vectors to get insights about the similarity between the corresponding datasets. Figure 4.6a and 4.6b present heatmaps of similarity between the debunking contexts and overall contexts of the datasets respectively. It is clear that the performance of the cross-domain models is strongly related to the dataset similarities. In particular, AraStance has the highest similarity with AuSTR on debunking context (0.20) and overall context (0.25) respectively. That resulted in the best performing cross-domain model achieving a macro-$F_1$ of 0.771 and $F_1$(*disagree*) of 0.687. Moreover, ArCOV19-Rumors has the second highest similarity with AuSTR on debunking context (0.10) and the second best performing cross-domain model achieving $F_1$(*disagree*) of 0.621. It is worth noting that although ArabicFC has the second highest similarity on the overall context, the model trained on it did not perform well especially on the *disagree* class, with $F_1$ of 0.332, due to the severe imbalance as mentioned earlier.

In summary, we found that AraStance is the best existing stance dataset for training a model for the task, as it covers a large number of fact-checked claims spanning multiple Arabic countries and topics compared to the other datasets.

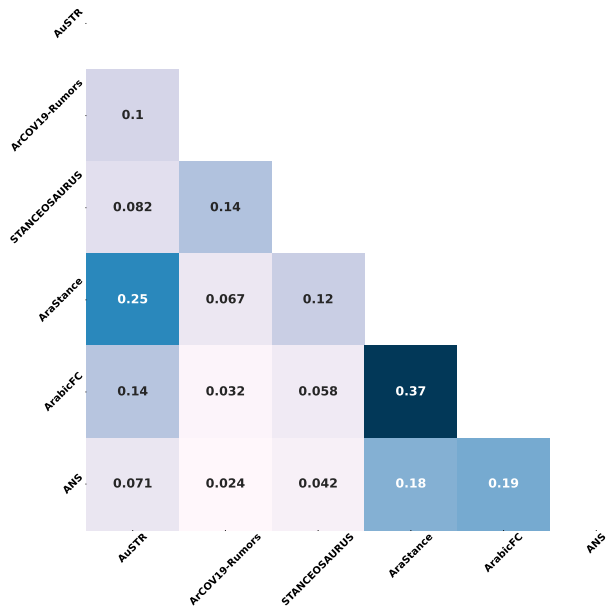
To answer **RQ5**, we conclude that some cross-domain stance datasets are somewhat useful for detecting the stance of authorities. However, motivated by the findings of Ng and Carley [210] who highlighted the potential benefit of aggregating datasets to enhance the stance detection, we were encouraged to conduct our subsequent experiments, in which we combine all cross-domain datasets for training.

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

Figure 4.6. Dataset pairwise similarity using (a) debunking contexts, and (b) overall contexts.




(a)




(b)

### *4.6.2. Combining Cross-domain Datasets for Training (RQ6)*

To address **RQ6**, we combined all cross-domain datasets and adopted the same setup mentioned previously, where we tune and test on AuSTR folds.

As presented in Figure 4.5, we note that, overall, the combined model achieved a *very-slightly* better performance in terms of macro-$F_1$ over the best individual model, i.e., the model trained with AraStance only. However, considering the individual classes, it exhibited the best performance for the *agree* class with a big margin compared to AraStance model; but it fell short for the *disagree* class. We speculate the reason is that some of the datasets, namely ANS and ArabicFC, achieved low performance on the *disagree* class, thus when combined with other datasets it affected negatively the overall performance on the same class.

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

To answer **RQ6**, we found that combining all cross-domain datasets can slightly improve the overall performance compared to the best performing individual model (AraStance), but could not beat it on detecting debunking tweets.

### *4.6.3. Introducing In-domain Data for Training (RQ7)*

To address **RQ7**, we first trained a stance model with in-domain data only, i.e., AuSTR. We then trained a model with in-domain data augmented with each of the cross-domain datasets separately and also with all cross-domain datasets combined.

As expected, the model trained with AuSTR only outperforms all models trained with cross-domain datasets across all evaluation measures, as shown in Figure 4.5. More specifically, it outperforms their best (i.e., the model trained with AraStance) by 15.3%, 7.1%, and 7.9% in $F_1$(*disagree*), $F_1$(*agree*), and macro-$F_1$ respectively, showing a clear need to in-domain data.

What if we augment AuSTR with the cross-domain datasets in training? Figure 4.7 illustrates that effect. For every single cross-domain dataset, when augmented with AuSTR, the resulted model outperforms the model trained only on the cross-domain data by a big margin, ranging from 6.8% to 35.6% in macro-$F_1$. This re-emphasizes the effect of in-domain data. However, only the model trained on AuSTR+AraStance was able to outperform the AuSTR-only model in macro-$F_1$ and $F_1$(*agree*) but not $F_1$(*disagree*). It turned out that augmenting AuSTR with AraStance made the *disagree* class minority, constituting only 13.3% of the training examples compared to 24.3% of AuSTR training examples, which negatively affects the performance on that class.

Contrary to the results presented in Figure 4.5, augmenting AuSTR with all cross-domain datasets achieved the lowest macro-$F_1$ compared to augmenting AuSTR with individual cross-domain datasets. In fact, the combined training data becomes clearly dominated with the cross-domain data (24,313 vs. 811 examples), which leads to negligible effect of the in-domain data.

To answer **RQ7**, we conclude that in-domain data is needed for better detecting the stance of authorities. Moreover, augmenting AuSTR with AraStance improved

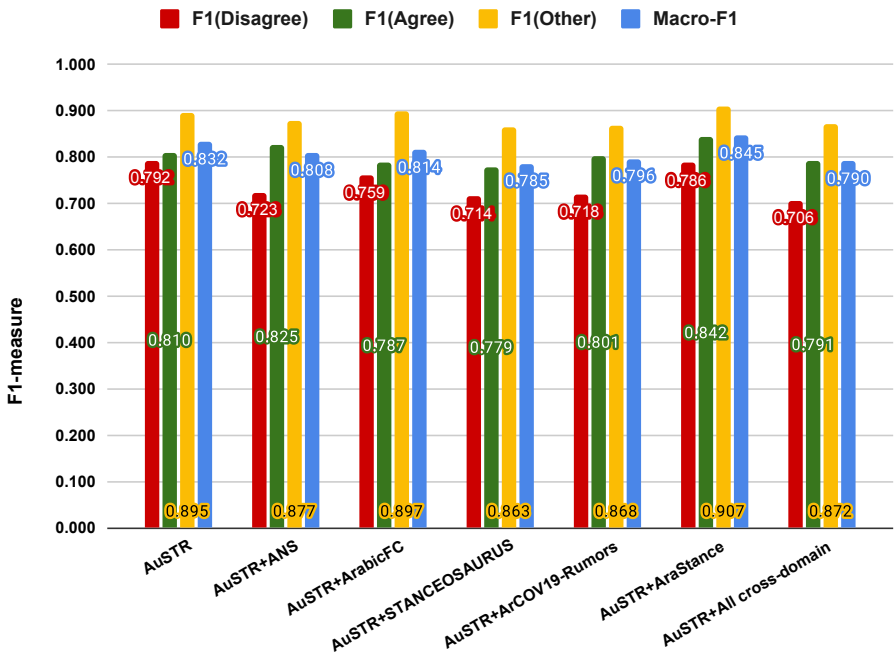



Figure 4.7. Performance of models trained using *in-domain* vs. *in-domain-augmented* data.

the overall performance but at the expense of degrading the performance on detecting debunking tweets, which, we argue, is more crucial for the task.

### *4.6.4. Addressing the Class-Imbalance Problem (RQ8)*

To address **RQ8**, we selected the best two models presented in Figure 4.7, namely the one trained with AuSTR only and the one trained with AuSTR augmented with AraStance. We then fined-tuned the stance models with the same previous setup but with two other loss functions, $WCE$ and $CBF$, as described in Section 4.5.5.

As presented in Table 4.6, we observe that adopting $WCE$ loss function could not improve the performance of the models compared to adopting $CE$. However, for the model trained with AuSTR, adopting $CBF$ notably improved the performance over $CE$ with about 4.2% on the *agree* class, which is the minority class in AuSTR data. However, it slightly degraded the performance on the *disagree* class. Overall, it improved macro-$F_1$ performance getting it closer to the performance of the model trained on AuSTR augmented with AraStance (0.843 vs. 0.845).

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

To answer **RQ8**, we conclude that adopting $CBF$ in addition to training on AuSTR solely is on bar with the model trained on both AuSTR and AraStance, nullifying the need for augmenting AuSTR with any cross-domain data for training.

Table 4.6. Training with different loss functions. Boldfaced and underlined numbers are the best and second best respectively per measure.

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

## 4.7. Discussion

In this section, we discuss our evaluation results in terms of failure cases (Section 4.7.1) and limitations (Section 4.7.2).

### *4.7.1. Failure Analysis*

We conducted a detailed error analysis on the 113 examples (constituting 14% of the data) that failed to be predicted correctly by the model trained with AuSTR and adopting $CBF$ loss. We categorize the reasons behind these errors based on a thorough examination of the failed pairs. We found that the failures can be attributed to six main reasons which we discuss below. Some failed examples are presented in Table 4.7.

1. **Implicit stance:** When an authority indirectly *agree* or *disagree* with the rumor. For example, $P_1$ is an example of a rumor about the infection of Mahmoud Al-Khatib, the director of Al-Ahly Egyptian football club, with COVID19, and an authority tweet implicitly debunking the rumor mentioning that he is attending the training session of the team in the stadium. This failure type is the cause of 30.09% of all failures, which motivates the need to address this challenge using stance models that take this into consideration.

2. **Writing style:** Where an authority is speaking about herself, e.g., $P_2$. Based on our examination, 12.39% of the failures are due to this reason.

3. **Misleading debunking keywords:** when an authority is either debunking another rumor that is relevant to the topic of the target rumor, or just including some debunking keywords in his tweets even when supporting a rumor. For example, in $P_3$, the authority tweet mentions that the "information being posted on it today is false.", although it is *agreeing* with the rumor. We found that this constitutes 10.62% of the failures.

4. **Misleading relevant keywords:** when an authority post tweets relevant to the topic of the rumor, the model may fail to predict the stance correctly, e.g., in $P_4$. This constitutes 25.66% of the failed examples.

5. **Lack of context**: when an authority debunks or supports a rumor by an announcement embedded in an image or a video, e.g, in $P_5$. This motivates the need to

consider the tweet multi-modality [29], [214] at the processing step. Moreover, some rumors may need additional context in order to be considered relevant to the authority tweet. We observed that 6.19% of the failures are of this type.

6. **Arabic MSA by authorities vs. dialects by normal users:** As opposed to English, working with Arabic language is very challenging as different dialects, i.e., informal languages, are used in different Arabic countries [215]. These dialects may have different vocabulary than the Modern Standard Arabic (MSA) which is usually used in formal communications [216]. Authority tweets are usually in formal language and written in MSA Arabic, while normal users may use their informal Arabic with variant dialects, which make detecting the stance more challenging.

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

Table 4.7. Sample examples failed to be predicted correctly by our best model. Failure types are *implicit stance*, *writing style*, *misleading debunking keywords*, *misleading relevant keywords*, and *lack of context* in order.

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

### *4.7.2. Limitations of Our Study*

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

#### *4.7.2.1. Data*

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

#### *4.7.2.2. Stance Models*

In our work, we adopted a BERT-based stance model, but we did not experiment with other models, e.g., [217] which might improve the performance we achieved. Moreover, we only experimented with ARBERT [168] as it showed to perform well for Arabic stance detection on most of our adopted cross-domain datasets [203]; however, we did not experiment with other Arabic BERT models [218].

## 4.8. Conclusion

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

In this chapter, we addressed detecting the stance of authorities towards rumors. In the next chapter, we target evidence retrieval from authority timelines where we investigate the usefulness of detecting the stance of authorities towards rumors for evidence retrieval from their timelines.

# CHAPTER 5: EVIDENCE RETRIEVAL FROM AUTHORITIES

In Chapter 4, we introduced incorporating authorities for rumor verification by exploiting the stance of their timeline tweets as a signal. In this chapter, we introduce *evidence retrieval from authorities*. Specifically, we propose retrieving evidence tweets from the authority timelines to be utilized as another source of evidence for rumor verification. Moreover, we investigate the usefulness of detecting the stance of authorities towards rumors for evidence retrieval from their timelines.

Several research studies addressed rumor verification in social media by mainly incorporating the propagation networks as a source of evidence. They either utilized the stance of the replies [15], [17], [219], structure of replies [6], [10], [220], or user metadata [18]. Recently, evidence from the Web was proposed to further augment signals from the conversation threads [19], [75].

Authorities (i.e., entities having the real knowledge or power to verify or deny a specific rumor) 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. A closer look to the literature on rumor verification in social media reveals that there is no study to date exploring incorporating *evidence tweets* retrieved from the timelines of authorities for rumor verification in social media. Moreover, there is no available dataset for that task to support such research.

This chapter starts with defining the task of *evidence retrieval from authorities* in Section 5.1. We present an overview of our work in Section 5.2. In Section 5.3, we discuss our dataset construction approach, analysis, and annotation challenges. Our experimental design and setup are presented in Sections 5.4 and 5.5 respectively. We analyze our experimental results and answer our research questions in Section 5.6. We present the limitations of this work in Section 5.7. Finally, we conclude in Section 5.8.

## 5.1. Problem Definition

We propose the task of *evidence retrieval from authorities* 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, retrieve the top N evidence tweets from those timelines.

## 5.2. Overview of Our Work

To facilitate the research on our proposed task, we introduce and publicly release AuRED, the first Authority-Rumor-Evidence Dataset. AuRED covers 160 *Arabic* rumors annotated with tweet-level evidence from their corresponding 692 authority timelines, comprising about 34k annotated tweets in total.

We consider our problem as a special case of the evidence retrieval for *fact-checking* problem [78], where evidence sentences are retrieved from relevant Wikipedia pages to verify a given claim. In our work, we assume that authorities for a given rumor are already retrieved (refer to Chapter 3), hence we only focus on *evidence retrieval* from the authority Twitter accounts. To that end, we further support the research on the new problem by providing benchmarking performance results of strong baseline models that were previously proposed for evidence retrieval for fact-checking. Our contributions are as follows:

1. We introduce the new task of *evidence retrieval from authorities over Twitter*.

2. We construct and release AuRED,[6] the first *Arabic* public dataset for the task.

3. We present benchmarking results on AuRED, and release our source code for reproducibility and facilitating research on the task.

4. We explore how existing evidence retrieval models proposed for fact-checking perform on our task, and if existing fact-checking datasets have the knowledge transfer potential to our task.

5. We also investigate the usefulness of detecting stance of authorities toward rumors for the evidence retrieval task.

## 5.3. 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 [21], yet under-explored for rumor verification. As presented in Figure 5.1, the dataset was constructed by annotating a set of rumors, selected from two existing datasets (Section 5.3.1) following two main steps (1) finding authorities that can help verify the rumors (Section 5.3.2), and (2) collecting and annotating the timelines of those authorities to find evidence tweets (Section 5.3.3).

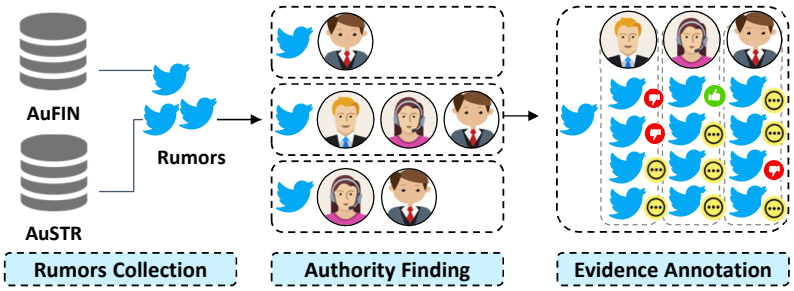



Figure 5.1. AuRED construction process.

### *5.3.1. Rumors Collection*

Due to time and budget constraints, we randomly selected 160 rumors from AuFIN and AuSTR datasets introduced in Chapter 3 and Chapter 4 respectively. 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. Thus, 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.

### 5.3.2. Authority Finding

AuFIN rumors are associated with their relevant authorities, however AuSTR rumors are only associated with an authority tweet either supporting, refuting or irrelevant to the rumor. Thus, for AuSTR rumors, in addition to considering the authority of the associated authority tweet, we collected more authorities for each rumor in the dataset following the same approach adopted to construct AuFIN dataset (refer to Section 3.4). 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.

### 5.3.3. 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 timelines.[1] 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 [158] as 0.67, which indicates "substantial" agreement [159]. 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. We present some example rumors and corresponding evidence tweets from our dataset, and its statistics in Table 5.1 and Table 5.2 respectively.

### 5.3.4. 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

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

Table 5.1. Sample rumors and corresponding authority evidence tweets (their English translation) from AuRED. The refuted and supported rumors have more than one evidence, but only one is presented for demonstration purposes.

| |
|---|
| **Refuted Rumor:** Moroccan reports: Bakary Gassama, is the referee of the return match between Al-Ahly and Wydad #195Sports [Link] [21-10-2020]<br>**Authority Evidence:** [@AlAhlyTV] Learn about the biography of referee Gomez, referee of the Al-Ahly and Wydad match today YouTube: [Link] #Six #Africa_Ahly #Alahlytv [23-10-2020]<br>**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 [Link] [22-12-2022]<br>**Authority Evidence:** [@Mofa_Libya] The account has been officially restored. We thank everyone who contributed and cooperated with us. @GovernmentLY @Hakomitna [21-12-2022]<br>**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 [Link] @marsdnews24 [05-01-2021]<br>**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] |

had been already fact-checked by our sources (e.g., rumor (b) is verified by "Misbar" fact-checking platform[2]), assuming those are viral, consequently could have higher impact on the community.

**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 authorities' 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**: Evidences 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 5.2a 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 which denies rumor (a). On the other hand, Figure 5.2b shows a video of the health minister

[2] http://tinyurl.com/496vysc5

Table 5.2. 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 | 33705 | |
| Average per rumor | 210.66 | |
| SUPPORTS | 118 | |
| REFUTES | 306 | |
| NOT ENOUGH INFO | 33281 | |
| Videos | 4998 | |
| Images | 17817 | |
| **AuRED* Authority tweets** | | |
| Authorities | 160 | |
| Average per rumor | 1 | |
| Authority tweets | 9755 | |
| Average per rumor | 60.97 | |
| SUPPORTS | 75 | |
| REFUTES | 213 | |
| NOT ENOUGH INFO | 9467 | |

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 text which requires greater effort.

### *5.3.5. Dataset 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 5.3 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 [221] 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.
**Domain Coverage**: We define *domain* here as the topic of the rumor such as politics, health, sports, etc. Figure 5.4 shows the diverse coverage of domains of rumors in AuRED.
**Multimodality**: To support the development of versatile verification systems, AuRED

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

Figure 5.2. Multimodality of evidences in AuRED.

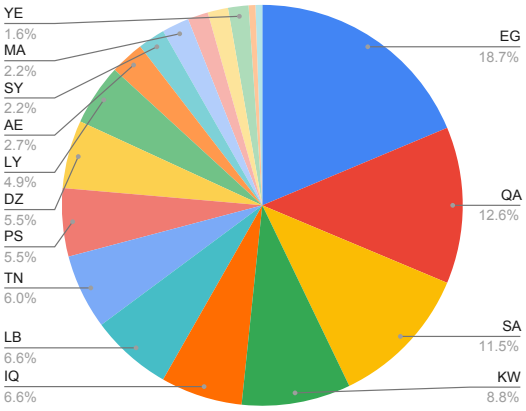



Figure 5.3. Geographical coverage of rumors in AuRED. The countries are represented by their 2-letter ISO codes.

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.

## 5.4. Experimental Design

Our task is closely related to the evidence retrieval for *fact-checking* [78]. 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 [222], it is intriguing to investigate how existing evidence retrieval for fact-checking 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 (e.g., FEVER [78], an English fact-checking dataset containing claims and their relevant evidence sentences extracted from Wikipedia pages), it is then intuitive to explore the potential of cross-lingual transfer learning. Accordingly, we address the following research questions:

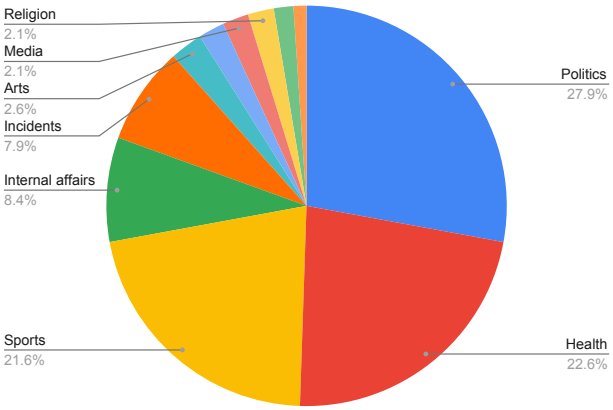



Figure 5.4. Domain coverage of rumors in AuRED.

- **RQ9**: How effective are existing evidence retrieval models for fact-checking for our task under the *cross-lingual zero-shot* setup? (Section 5.6.1)
- **RQ10**: How do existing evidence retrieval models for fact-checking models perform on our task if they are directly *fine-tuned* with AuRED? (Section 5.6.2)

To address both research 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, 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.

## 5.5. Experimental Setup

In this section, we present our detailed experimental setup. We discuss our adopted evidence retrieval models in Section 5.5.1. We also discuss how we evaluate those models for our task in Section 5.5.2.

### *5.5.1. Evidence Retrieval Models*

In addition to evaluating strong sparse and dense retrieval approaches, we selected two SOTA models for evidence retrieval which exhibited the best performance on FEVER test set Park, Lee, Jeon, *et al.* [223].[3] 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.

- **BM25**: One of the most successful lexical retrieval models [160]. Using Pyserini [186], we constructed an index for all tweets from all authorities for a given

[3]The model proposed recently by DeHaven and Scott [88] is SOTA but they adopt re-retrieval using hyperlinks in retrieved sentences to beat KGAT. Re-retrieval is not applicable in our work

rumor. We then retrieved, for each rumor, the top 5 relevant authority tweets from the corresponding index.

- **mContriever** [224]: It is a multilingual dense retrieval model that achieves good retrieval performance on Arabic data when further fine-tuned using MS MARCO dataset [224]. We retrieved the top 5 tweets (out of the corresponding authority tweets) that are the closest in the Contriever's embedding space using cosine similarity.

- **KGAT** [93]: It is a widely adopted retrieval model in fact-checking studies [84], [223], [225], [226]. 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 scores predicted for each rumor-tweet-authority-tweet pair is used to retrieve the top 5 evidence tweets. We adopted the authors' implementation.[4]

- **MLA** [112]: It is a pointwise BERT-based binary classifier to detect evidence vs. non-evidence. The cross entropy loss was adopted. The authors proposed sampling M non-evidence sentences from the labelled evidence set 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.[5] At training and inference, rumor-tweet-authority-tweet pair are fed to a BERT-based model separated by a [SEP] token. The authors' source code was adopted for our experiments.[6]

- **TML [227]:** As explored for evidence retrieval for fact-checking by Bekoulis, Papagiannopoulou, and Deligiannis [227], we investigate the performance when adopting the triplet 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' source code.[7]

- **STAuRED**: Motivated by the task of detecting the stance of authorities, introduced in Chapter 4, 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 text as sentence A and the authority tweet text 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

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

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

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

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

three output nodes, added on top of BERT architecture, to compute the probability for each stance class. For retrieving the top 5 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 [168],[8] 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 [78], but we replaced the English BERT with multilingual BERT (mBERT) [228].[9]

#### *5.5.2. Evaluation Scenarios and Measures*

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 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). Additionally, 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 [78], 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.

### 5.6. Results and Discussion

In this section, we present and discuss the results of our experiments which address the two research questions introduced in Section 5.4.

#### *5.6.1. Cross-lingual Zero-shot Scenario (RQ9)*

For this setup, we fine-tuned MLA and KGAT evidence retrieval models presented in Section 5.5.1 using the authors' setup. Since, for this scenario, we train on

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

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

Table 5.3. Performance of Cross-lingual Zero-shot Evidence Retrieval. Bold scores are the best for each test set.

| Test Set | Retrieval Model | **Standard IR Scores** | | **FEVER Scores** | | |
|---|---|---|---|---|---|---|
| | | MAP | R@5 | P@5 | R@5 | F1@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 |

English data (FEVER) and test on Arabic data (AuRED), we adopted the multi-lingual BERT (mBERT) as the pre-trained model. The models were then used to retrieve evidence for AuRED test rumors. We report the average performance using cross-validation in Table 5.3.

As shown in Table 5.3, 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 due to the fact that evidence is retrieved from the timeline of a single authority for each rumor.

### *5.6.2. In-domain Fine-tuning Scenario (RQ10)*

For this setup, we tested the evidence retrieval models presented in Section 5.5.1. We fine-tuned all the models using AuRED, adopting a cross-validation setup. The performance of the models is presented in Table 5.4.

MLA and STAuRED are the two 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 human fact-checkers who are usually interested in more evidence to reach a solid verification decision.

Table 5.4. Performance of In-domain Fine-tuning for Evidence Retrieval. Bold and underlined scores are the best and second-best respectively for each test set.

| Test Set | Retrieval Model | **Standard IR Scores** | | **FEVER Scores** | | |
|---|---|---|---|---|---|---|
| | | MAP | R@5 | P@5 | R@5 | F1@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 |

## 5.7. Limitations of Our Study

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 [75], [229] 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.

## 5.8. Conclusion

In this chapter, we introduced the problem of *evidence retrieval from authorities*. 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 evidence retrieval for fact-checking models to set up the baseline systems for our task. Our experiments show that evidence retrieval models for fact-checking achieved competitive benchmark results even under cross-lingual zero-shot setup. For future work, we plan to (1) consider the multimodality of evidence tweets to improve the evidence retrieval, (2) propose models to improve the performance achieved, and (3) construct and release a similar dataset in English to further facilitate and encourage research on the task. In this chapter we addressed evidence retrieval from authorities problem. We study how to use those retrieved evidences for rumor verification in the next chapter.

# CHAPTER 6: RUMOR VERIFICATION USING EVIDENCE FROM AUTHORITIES

In chapter 5, we focused on evidence retrieval from authorities. In this chapter, we study how to use those retrieved evidences for rumor verification. Specifically, we propose *rumor verification using evidence retrieved from authorities* problem.

A few recent works in literature proposed incorporating evidence retrieved from Web articles for rumor verification in social media [19], [75]. 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.

This chapter starts with defining the task of *rumor verification using evidence from authorities* in Section 6.1. We present an overview of our work in Section 6.2. We present our experimental design and setup in Sections 6.3 and 6.4 respectively. We analyze our experimental results and answer our research questions in Section 6.5. We discuss the limitations of this study in Section 6.6. Finally, we conclude in Section 6.7.

## 6.1. Problem Definition

We propose the task of *rumor verification using evidence from authorities* defined as follows: Given a rumor expressed in a tweet and a set of retrieved evidence tweets from authority Twitter accounts for that rumor, the system should determine if the rumor is supported, refuted, or unverifiable.

## 6.2. Overview of Our Work

In our work, we consider the *rumor verification using evidence from authorities* a special case of the claim verification for *fact-checking* problem [78], where evidence sentences retrieved from relevant Wikipedia pages are used to verify a given claim. We assume that authorities for a given rumor and their evidence tweets are already retrieved (refer to Chapter 3 and 5), hence we only focus on *rumor verification* using the retrieved evidence from relevant authorities. We support the research on the new problem by providing benchmarking performance results of strong baseline models that were previously proposed for claim verification for fact-checking. The contributions of this chapter are as follows:

1. We introduce the new task of *rumor verification using evidence from authorities over Twitter*.

2. We present benchmarking results on AuRED, and release our source code for reproducibility and facilitating research on the task.

3. We explore how existing claim verification models proposed for fact-checking perform on our task, and if existing fact-checking datasets have the knowledge transfer potential to our task.

## 6.3. Experimental Design

As mentioned previously, our task is closely related to the general task of *fact-checking* [78]. 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 [222], it is intriguing to investigate how existing 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 (e.g., FEVER [78], an English fact-checking dataset containing claims and their relevant evidence sentences extracted from Wikipedia pages), it is then intuitive to explore the potential of cross-lingual transfer learning. Accordingly, we address the following research questions:

- **RQ11**: How effective are existing claim verification for fact-checking models for our task under the *cross-lingual zero-shot* setup? (Section 6.5.1)
- **RQ12**: How do existing claim verification for fact-checking models perform on our task if they are directly *fine-tuned* with AuRED? (Section 6.5.2)

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 claim 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.

## 6.4. Experimental Setup

In this section, we present our detailed experimental setup. We discuss our adopted rumor verification models in Section 6.4.1. We also discuss how we evaluate those models for our task in Section 6.4.2.

### *6.4.1. 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:[1]

1. **MLA [112]:** It adopts multi-task learning considering the claim verification as the main task, and evidence retrieval as an auxiliary task where it incorporates the evidence retrieval scores through joint training. The model takes as input a claim and 5 evidence sentences, and represent each evidence using BERT [CLS] token. It then applies token-level attention over a claim-evidence pair, token and sentence-level self-attentions for evidence sentences. Finally, it combines all hidden states together with the evidence retrieval scores at the final attention layer.
2. **KGAT [93]:** It is 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.

[1]The model proposed recently by DeHaven and Scott [88] 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 as mentioned previously due to the re-retrieval step which is not applicable to our task.

Table 6.1. Performance of Cross-lingual Zero-shot Rumor Verification. Bold scores are the best for each test set.

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

**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 for each fold.

### *6.4.2. Evaluation Measures*

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 [78], we also consider *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 for a specific rumor.

## 6.5. Results and Discussion

In this section, we present and discuss the results of our experiments which address the two research questions introduced in Section 6.3.

### *6.5.1. Cross-lingual Zero-shot Scenario (RQ11)*

For this setup, we fine-tuned MLA and KGAT models presented in Section 6.4.1 using the authors' setup for the claim verification task. Since, for this scenario, we train on English data (FEVER) and test on Arabic data (AuRED), we adopted the multi-lingual BERT (mBERT) as the pre-trained model. The models were then used to verify AuRED test rumors using the retrieved evidence by MLA and KGAT retrieval models (presented in Section 5.6.1). We report the average performance, using cross-validation, for rumor verification in Table 6.1.

As presented in Table 6.1, 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, as presented in Table 5.1. Recall that FEVER claims are generated by manipulating the Wikipedia sentences adopting paraphrasing, negation, or entity substitution to name a few changes [230]. Thus, the models may have learned a different styles of evidence to decide whether a given a rumor REFUTES, SUPPORTS, or NOT ENOUGH INFO to verify it. Finally, we observe that

Table 6.2. Performance of In-domain Fine-tuning for Rumor Verification. Bold scores are the best for each test set.

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

KGAT significantly outperforms MLA in rumor verification, despite the superiority of the latter in evidence retrieval, showing clearly that the retrieval and verification models are different.

### *6.5.2. In-domain Fine-tuning Scenario (RQ12)*

For this setup, we fine-tuned all the models using AuRED, adopting a cross-validation setup. The performance of rumor verification is presented in Table 6.2.

Neither of the models perform well on this task, indicating there is 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 [231] or seq2seq text generation models [232], or *domain adaptation* [233] over fact-checking datasets. While KGAT still exhibits better performance than MLA when fine-tuned with in-domain training data, the performance interestingly has not reached the performance under the cross-lingual setup shown in Table 6.1. This can be attributed to the size of the training data in both cases; the big collection of claims in FEVER (having 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.

## 6.6. Limitations of Our Study

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

1. **Dataset size**: The small number of rumors in our AuRED dataset (160 rumors), 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.

2. **Verification models**: In our work, we experimented with only two existing claim verification for fact-checking models. We believe further exploration to experiment with more models will lead to better findings and conclusions.

3. **Extrinsic evaluation**: In all our rumor verification experiments, we considered evidence tweets retrieved by the retrieval models presented in Chapter 5. However, we did not perform any intrinsic evaluation where we consider gold evidence tweets (decided by the annotators) for rumor verification.

## 6.7. Conclusion

In this chapter, we introduced the problem of *rumor verification using evidence from authorities over Twitter*. We investigated the effectiveness of existing claim verification for fact-checking models to set up the baseline systems for our task. Our experiments showed that performance of rumor verification is still far from enough. For future work, we plan to augment the dataset to expand the number of rumors to improve the rumor veracity prediction, and further enhance the models to improve the performance achieved.

# CHAPTER 7: CONCLUSION AND FUTURE WORK

In this chapter, we conclude with a summary of the results of this dissertation in Section 7.1. We then thoroughly discuss the implications of this dissertation in Section 7.2. The limitations of this work are presented in Section 7.3. Finally, we present the future directions of our work in Section 7.4.

## 7.1. Conclusion

This dissertation contributes towards a crucial task, i.e., rumor verification in social media. We proposed augmenting the traditional rumor verification pipeline by incorporating authorities as another source of evidence. Specifically, in this dissertation we introduced the problem of *rumor verification using evidence from authorities* which we decomposed into a pipeline of four sub-problems namely 1) *authority finding in Twitter*, 2) *detecting the stance of authorities towards rumors in Twitter*, 3) *evidence retrieval from authorities* and 4) *rumor verification using evidence from authorities*. To address the problem we constructed and released three Arabic datasets namely 1) the first Authority FINding in Twitter (AuFIN), 2) the first Authority STance towards Rumors (AuSTR), 3) the first Authority-Rumor-Evidence Dataset (AuRED). We addressed each sub-problem and introduced competitive baseline models. Drawing upon our experiments, we discussed failure factors and presented recommendations for future research directions in addressing each task.

As for the *authority finding in Twitter* task, we introduced AuFIN dataset which comprises 150 rumors (expressed in tweets) associated with a total of 1,044 authority accounts and a user collection of 395,231 Twitter accounts (members of 1,192,284 unique Twitter lists). Moreover, we proposed a hybrid model that employs pre-trained language models and combines lexical, semantic, and network signals to find authorities. Our experiments showed that the textual representation of users is insufficient, and incorporating the Twitter network features improved the recall of authorities by 34%. Moreover, semantic ranking is inferior to the lexical and network-based ranking in terms of precision, but superior in terms of recall. Therefore, combining both the semantic and network-based ranking achieved the best overall performance achieving a precision of 0.413 and 0.213 at depth 1 and 5 respectively. We showed that rumor expansion by exploiting Knowledge Bases improves the recall of authorities by up to 15%. Furthermore, we found that SOTA models for topic expert finding perform poorly on finding authorities. Finally, our *Authority Finding* model was deployed as part Tahaqqaq, a real-time system for assisting Twitter users in Arabic claim verification, to enable users to find authorities for a given tweet in real-time or any free-text claim.

As for *detecting the stance of authorities towards rumors in Twitter*, we constructed AuSTR dataset which comprises 811 (rumor tweet, authority tweet) pairs relevant to 292 unique rumors. Due to the relatively limited size of our dataset, we explored the adequacy of existing Arabic datasets of stance towards claims in training BERT-based models for our task, and the effect of augmenting AuSTR with those datasets. Our experiments showed that, despite its limited size, a model trained solely on AuSTR with a class-balanced focal loss exhibits a comparable performance to the best studied combination of existing datasets augmented with AuSTR, achieving a performance of 0.84 macro-F1 and 0.78 F1 on debunking tweets. The results indicated that AuSTR can be sufficient for our task without the need for augmenting it with existing stance datasets.

As for the *evidence retrieval from authorities*, we introduced AuRED which

comprises 160 rumors expressed in tweets and 692 Twitter timelines of authorities comprising about 34k annotated tweets in total. We explored how existing evidence retrieval for fact-checking models perform on our task, and if existing fact-checking datasets have the knowledge transfer potential to our task. Our experiments showed that although evidence retrieval models perform relatively well on the task establishing strong baselines, 0.70 as recall at depth 5, there is still a big room for improvement. The results also showed 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 setups and datasets is required.

As for the *rumor verification using evidence from authorities*, we investigated how existing claim verification for fact-checking models perform on our task, and if existing fact-checking datasets have the knowledge transfer potential to our task. Our experiments showed that due to the small size of the dataset (160 rumors), the performance on rumor verification is still far from enough achieving a performance of 0.422 macro-F1.

Finally, we co-organized three out of the four sub-problems in our proposed *rumor verification using evidence from authorities* pipeline as shared tasks in CLEF 2023, and CLEF 2024 CheckThat! labs to motivate the research community to work on our problem.

## 7.2. Research Implications

In this section, we discuss the theoretical and practical implications of our work.

### *7.2.1. Theoretical Implications*

A myriad of systems for rumor verification in social media were proposed in the literature, but they mainly relied on subjective evidence, e.g., propagation networks or user interactions. However, there is no work that addressed exploiting evidence from trusted authorities. To fill this gap, we presented the first study for rumor verification using evidence from authorities over Twitter. The theoretical implications of our study are as follows:

- **Introducing new research problems with clear definitions:** We introduced and defined a pipeline of four research problems namely 1) *authority finding in Twitter*, 2) *detecting the stance of authorities towards rumors in Twitter*, 3) *evidence retrieval from authorities*, and 4) *rumor verification using evidence from authorities*.

- **Bringing attention to the rumor verification using evidence from authorities over Twitter problem:** We motivated and defined rumor verification using evidence from authorities over Twitter problem, which we believe can help fact checkers and automated rumor verification systems to find the right authorities and evidence from their Twitter timelines, hence helping in the verification process.

- **Encouraging the construction of datasets to address the problem in other languages:** We constructed and released 1) the first Authority FINding in Twitter (AuFIN), 2) the first Authority STance towards Rumors (AuSTR), 3) the first

Authority-Rumor-Evidence Dataset (AuRED), along with our construction approach and annotation guidelines to encourage the construction of similar datasets in other languages.

- **Proposing future directions:** Our presented error analysis, discussed issues, and limitations reveal a proposed roadmap for some future directions in this problem.

#### *7.2.2. Practical Implications*

Our proposed system can be exploited by fact-checkers or journalists, in addition to researchers targeting rumor verification in Twitter. Thus, we believe that the practical implications of our work are as follows:

- **Establishing baselines for further research:** Given that our best-performing models specifically authority finder and rumor verification achieved modest performance, this implies that the tasks (represented by our datasets) are indeed challenging and requires further development to achieve better performance. Although the performance of our models for detecting the stance of authorities and evidence retrieval from authorities was better compared to the other tasks there is still a room for improvements addressing the limitations discussed in the corresponding chapters (Chapter 4 and Chapter 5). More importantly, our approach establishes a strong baseline for future studies on this problem.

- **Building a system for journalists and fact-checkers:** Journalists or fact-checkers who attempt to verify a rumor over social media try to find a trusted source of evidence (relevant to that rumor) that can help them confirm or deny that specific rumor. A strong source of evidence for verifying a rumor is an *authority* who has the "real knowledge" to verify it if asked to. Our proposed work is a step towards providing such service.

- **Supporting automated verification systems:** Finding authorities and their evidence can be integrated into an automated verification pipeline for better rumor verification.

### 7.3. Limitations of Our Work

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

1. **Datasets:** As discussed previously, one of the limitations of this work is the size of our datasets especially AuRED dataset which comprises 160 rumors only. Moreover, in this work we only targeted the Arabic language.

2. **Models:** Our best-performing models specifically authority finder and rumor verification achieved modest performance, and requires further development to achieve better performance.

3. **Limited pre-processing:** Our models use the tweets mostly as is, with little pre-processing. We only used the textual content of the tweets without considering expanding the context of the rumor with extracted text from images, videos, or external news articles which may improve the retrieval of authorities and their evidence.

4. **Limited evaluation:** In our work, we did not consider the full pipeline evaluation. We evaluated evidence retrieval and rumor verification using evidence from authorities assuming that the authorities are already retrieved. Moreover, we proposed augmenting existing sources of evidence with authorities' evidence however we did not evaluate rumor verification with all existing evidence sources including our proposed one.

5. **Effectiveness only:** In our work we only targeted the effectiveness of our pipeline models. However, the efficiency and the scalability of our proposed pipeline was not addressed.

## 7.4. Future Directions

There are several directions for future work. We elaborate on some of them in the following:

1. **Datasets expansion and language coverage:** We believe expanding the dataset can improve the performance of the models, and it can lead to better generalizable models. Moreover, constructing and releasing similar datasets in English and other languages will further facilitate and encourage research on the problem.

2. **Evaluation:** Our evaluation to our proposed problem is limited in some aspects and can be addressed in future studies as discussed below:

    - **Full pipeline evaluation:** Although in our work we considered the retrieved evidences retrieved from our models for rumor verification, we assumed that the authorities are already retrieved. We plan to evaluate our full proposed pipeline by retrieving evidences from the retrieved authorities by our model.
    - **Extrinsic and intrinsic evaluation:** We plan to perform both *extrinsic*, i.e, predictions is done using the output of the preceding component, and *intrinsic*, i.e, predictions is done using the annotated data for both the evidence retrieval and rumor verification tasks. In our work, we performed *intrinsic* evaluation for the evidence retrieval (retrieve evidence from authorities decided by annotators) and *extrinsic* evaluation for rumor verification (evidences retrieved from our retrieval models).
    - **Augmenting existing sources of evidence:** Since we proposed augmenting existing rumor verification pipeline with evidence retrieval from authorities, we plan to augment our data with other sources of evidence and perform an ablation study for rumor verification using all the sources of evidences.

3. **Models Enhancement:** Based on our findings, we plan to further enhance our models. For authority finding, we plan to explore other sources for representing users, and exploiting other features to differentiate experts from authorities. Moreover, exploring methods for selecting relevant entities for rumor expansion may improve authorities retrieval. For detecting the stance of authorities, we plan to explore other stance models that can detect the implicit stance and take the authorities writing style into consideration. Furthermore, we plan to explore different approaches for evidence retrieval and rumor verification.

4. **Multimodality and context expansion:** Based on our failure analysis, we recommend considering the multimodality and context expansion of both the rumor and authorities tweets to address all the tasks in our proposed pipeline.

5. **Models deployment:** We plan to deploy our full proposed pipeline into a real-time rumor verification system as we did for our authority finding model [33].

6. **Efficiency and scalability:** In our work we focused on the effectiveness of the models. Given that early rumor verification is crucial, and because we are proposing a pipeline of multiple components, we plan to analyze the efficiency of our proposed pipeline when deployed as part of a real-time rumor verification system.

REFERENCES


[1] S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] B. Nyhan and J. Reifler, "Displacing misinformation about events: An experimental test of causal corrections," *Journal of experimental political science*, vol. 2, no. 1, pp. 81–93, 2015.

[3] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.

[4] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PloS one*, vol. 11, no. 3, e0150989, 2016.

[5] Y. Liu and Y.-F. B. Wu, "FNED: A Deep Network for Fake News Early Detection on Social Media," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 3, pp. 1–33, 2020.

[6] C. Song, K. Shu, and B. Wu, "Temporally evolving graph neural network for fake news detection," *Information Processing & Management*, vol. 58, no. 6, p. 102 712, 2021.

[7] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 72–81.

[8] S. Roy, M. Bhanu, S. Saxena, S. Dandapat, and J. Chandra, "GDART: Improving rumor verification in social media with Discrete Attention Representations," *Information Processing & Management*, vol. 59, no. 3, p. 102 927, 2022.

[9] J. Ma, W. Gao, and 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.

[10] T. Bian, X. Xiao, T. Xu, *et al.*, "Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 549–556.

[11] J. Choi, T. Ko, Y. Choi, H. Byun, and C.-k. Kim, "Dynamic graph convolutional networks with attention mechanism for rumor detection on social media," *Plos one*, vol. 16, no. 8, e0256039, 2021.

[12] N. Bai, F. Meng, X. Rui, and Z. Wang, "Rumor detection based on a Source-Replies conversation Tree Convolutional Neural Net," *Computing*, vol. 104, no. 5, pp. 1155–1171, 2022.

[13] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and 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)*, Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 69–76.

[14] L. Wu, Y. Rao, H. Jin, A. Nazir, and L. Sun, "Different absorption from the same sharing: Sifted multi-task learning for fake news detection," *arXiv preprint arXiv:1909.01720*, 2019.

[15] S. Kumar and K. Carley, "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*, Florence, Italy: Association for Computational Linguistics, Jul. 2019. DOI: 10.18653/v1/P19-1498.

[16] L. Chen, Z. Wei, J. Li, B. Zhou, Q. Zhang, and X.-J. Huang, "Modeling Evolution of Message Interaction for Rumor Resolution," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6377–6387.

[17] J. Yu, J. Jiang, L. M. S. Khoo, H. L. Chieu, and R. Xia, "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)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1392–1401. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.108.

[18] Y. Liu and Y.-F. B. Wu, "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[19] J. Dougrez-Lewis, E. Kochkina, M. Arana-Catania, M. Liakata, and 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.

[20] X. Hu, Z. Guo, J. Chen, L. Wen, and 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*, 2023, pp. 2901–2912.

[21] T. Alshaabi, D. R. Dewhurst, J. R. Minot, *et al.*, "The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020," *EPJ data science*, vol. 10, no. 1, p. 15, 2021.

[22] M. Hasanain, F. Haouari, R. Suwaileh, *et al.*, "Overview of CheckThat! 2020 Arabic: Automatic Identification and Verification of Claims in Social Media," in *CLEF*, 2020.

[23] M. K. Elhadad, K. F. Li, and F. Gebali, "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*, Springer, 2021, pp. 256–268.

[24] A. R. Mahlous and A. Al-Laith, "Fake News Detection in Arabic Tweets during the COVID-19 Pandemic," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

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

[26] S. Alqurashi, B. Hamoui, A. Alashaikh, A. Alhindi, and E. Alanazi, "Eating Garlic Prevents COVID-19 Infection: Detecting Misinformation on the Arabic Content of Twitter," *arXiv preprint arXiv:2101.05626*, 2021.

[27] A. Sawan, T. Thaher, and 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.

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

[29] R. M. Albalawi, A. T. Jamal, A. O. Khadidos, and A. M. Alhothali, "Multimodal Arabic Rumors Detection," *IEEE Access*, 2023.

[30] F. Haouari and T. Elsayed, "Are authorities denying or supporting? Detecting stance of authorities towards rumors in Twitter," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 34, 2024.

[31] F. Haouari, T. Elsayed, and W. Mansour, "Who can verify this? Finding authorities for rumor verification in Twitter," *Information Processing & Management*, vol. 60, no. 4, p. 103 366, 2023.

[32] F. Haouari and T. Elsayed, "Detecting Stance of Authorities towards Rumors in Arabic Tweets: A Preliminary Study," in *Proceedings of the 45th European Conference on Information Retrieval (ECIR'23)*, 2023.

[33] Z. Sheikh Ali, W. Mansour, F. Haouari, M. Hasanain, T. Elsayed, and A. Al-Ali, "Tahaqqaq: A Real-Time System for Assisting Twitter Users in Arabic Claim Verification," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[34] F. Haouari, "Evidence-Based Early Rumor Verification in Social Media," in *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, 2022, pp. 496–504.

[35] F. Haouari, M. Essam, and T. Elsayed, "bigIR at TREC 2020: Simple but Deep Retrieval of Passages and Documents.," in *TREC*, 2020.

[36] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 82–91.

[37] A. Barrón-Cedeño, F. Alam, A. Galassi, *et al.*, "Overview of the CLEF–2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Source," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2023, pp. 251–275.

[38] P. Nakov, G. Da San Martino, T. Elsayed, *et al.*, "Overview of the CLEF–2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, Springer, 2021, pp. 264–291.

[39] A. Barrón-Cedeno, T. Elsayed, P. Nakov, *et al.*, "Overview of Checkthat! 2020: Automatic Identification and Verification of Claims in Social Media," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2020, pp. 215–236.

[40] A. Barrón-Cedeño, F. Alam, T. Chakraborty, *et al.*, "The CLEF-2024 Checkthat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness," in *European Conference on Information Retrieval*, Springer, 2024, pp. 449–458.

[41] A. Barrón-Cedeño, F. Alam, T. Caselli, *et al.*, "The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority," in *European conference on information retrieval*, Springer, 2023, pp. 506–517.

[42] P. Nakov, G. Da San Martino, T. Elsayed, *et al.*, "The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News," in *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, Springer, 2021, pp. 639–649.

[43] A. Barrón-Cedeño, T. Elsayed, P. Nakov, *et al.*, "CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media," in *European Conference on Information Retrieval*, Springer, 2020, pp. 499–507.

[44] F. Haouari, Z. Sheikh Ali, and 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*, ser. CLEF '2023, Thessaloniki, Greece, 2023.

[45] F. Haouari, T. Elsayed, and R. Suwaileh, "Overview of the CLEF-2024 CheckThat! Lab Task 5 on Rumor Verification using Evidence from Authorities," in *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, G. Faggioli, N. Ferro, P. Galuščáková, and A. García Seco de Herrera, Eds., ser. CLEF 2024, Grenoble, France, 2024.

[46] S. Shaar, A. Nikolov, N. Babulkov, *et al.*, "Overview of CheckThat! 2020 English: Automatic Identification and Verification of Claims in Social Media," in *CLEF*, 2020.

[47] S. Shaar, M. Hasanain, B. Hamdan, *et al.*, "Overview of the CLEF-2021 CheckThat! Lab Task 1 on Check-Worthiness Estimation in Tweets and Political Debates," in *CLEF (Working Notes)*, 2021.

[48] S. Shaar, F. Haouari, W. Mansour, *et al.*, "Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates," in *CLEF (Working Notes)*, 2021.

[49] J. Ma, W. Gao, and K.-F. Wong, "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 708–717.

[50] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on Twitter," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 1867–1870. DOI: 10.1145/2806416.2806651.

[51] J. Ma, W. Gao, P. Mitra, *et al.*, "Detecting rumors from microblogs with recurrent neural networks," 2016.

[52] J. Ma, W. Gao, and K.-F. Wong, "Detect rumor and stance jointly by neural multi-task learning," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 585–593.

[53] Q. Li, Q. Zhang, and L. Si, "Rumor detection by exploiting user credibility information, attention and multi-task learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1173–1179.

[54] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[55] K. Radhakrishnan, T. Kanakagiri, S. Chakravarthy, and V. Balachandran, ""a little birdie told me..."-Social Media Rumor Detection," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 2020, pp. 244–248.

[56] A. Khandelwal, "Fine-tune longformer for jointly predicting rumor stance and veracity," in *8th ACM IKDD CODS and 26th COMAD*, 2021, pp. 10–19.

[57] G. Gorrell, E. Kochkina, M. Liakata, *et al.*, "SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 845–854.

[58] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and 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)*, Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 69–76. DOI: 10.18653/v1/S17-2006.

[59] Q. Huang, J. Yu, J. Wu, and B. Wang, "Heterogeneous Graph Attention Networks for Early Detection of Rumors on Twitter," *arXiv preprint arXiv:2006.05866*, 2020.

[60] Y. Geng, Z. Lin, P. Fu, and W. Wang, "Rumor detection on social media: A multi-view model using self-attention mechanism," in *International Conference on Computational Science*, Springer, 2019, pp. 339–352.

[61] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," in *2015 IEEE 31st International Conference on Data Engineering*, 2015, pp. 651–662. DOI: 10.1109/ICDE.2015.7113322.

[62] A. Dang, A. Moh'd, A. Islam, and E. Milios, "Early detection of rumor veracity in social media," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019. DOI: 10.24251/HICSS.2019.284.

[63] J. Ma, W. Gao, S. Joty, and K.-F. Wong, "An attention-based rumor detection model with tree-structured recursive neural networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 4, pp. 1–28, 2020. DOI: 10.18653/v1/P18-1184.

[64] C. Yuan, Q. Ma, W. Zhou, J. Han, and S. Hu, "Jointly embedding the local and global relations of heterogeneous graph for rumor detection," in *2019 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2019, pp. 796–805. DOI: 10.1109/ICDM.2019.00090.

[65] Y.-J. Lu and C.-T. Li, "GCAN: Graph-aware co-attention networks for explainable fake news detection on social media," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 505–514. DOI: 10.18653/v1/2020.acl-main.48.

[66] N. Rosenfeld, A. Szanto, and D. C. Parkes, "A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: Association for Computing Machinery, 2020, pp. 1018–1028, ISBN: 9781450370233. DOI: 10.1145/3366423.3380180. [Online]. Available: https://doi.org/10.1145/3366423.3380180.

[67] T. Sun, Z. Qian, S. Dong, P. Li, and Q. Zhu, "Rumor detection on social media with graph adversarial contrastive learning," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22, , Virtual Event, Lyon, France, Association for Computing Machinery, 2022, pp. 2789–2797, ISBN: 9781450390965. DOI: 10.1145/3485447.3511999. [Online]. Available: https://doi.org/10.1145/3485447.3511999.

[68] C. Song, Y. Teng, Y. Zhu, S. Wei, and B. Wu, "Dynamic graph neural network for fake news detection," *Neurocomputing*, vol. 505, pp. 362–374, 2022.

[69] J. Lin, R. Nogueira, and A. Yates, *Pretrained transformers for text ranking: Bert and beyond*, 2021. arXiv: 2010.06467 [cs.IR].

[70] J. Ma and W. Gao, "Debunking rumors on Twitter with tree transformer," ACL, 2020.

[71] L. Tian, X. Zhang, Y. Wang, and H. Liu, "Early detection of rumours on Twitter via stance transfer learning," in *European Conference on Information Retrieval*, Springer, 2020, pp. 575–588. DOI: https://doi.org/10.1007/978-3-030-45439-5_38.

[72] L. M. S. Khoo, H. L. Chieu, Z. Qian, and J. Jiang, "Interpretable rumor detection in microblogs by attending to user interactions," *arXiv preprint arXiv:2001.10667*, 2020.

[73] L. Alsudias and P. Rayson, "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*, K. Verspoor, K. B. Cohen, M. Dredze, *et al.*, Eds., Online: Association for Computational Linguistics, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.16.

[74] M. S. H. Ameur and H. Aliane, "AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset," *Procedia Computer Science*, vol. 189, pp. 232–241, 2021.

[75] X. Hu, Z. Guo, J. Chen, L. Wen, and 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*, ser. SIGIR '23, New York, NY, USA: Association for Computing Machinery, 2023, pp. 2901–2912.

[76] D. S. Nielsen and R. McConville, "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*, 2022, pp. 3141–3153.

[77] A. E. Lillie, E. R. Middelboe, and L. Derczynski, "Joint rumour stance and veracity prediction," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, M. Hartmann and B. Plank, Eds., Turku, Finland: Linköping University Electronic Press, Sep. 2019, pp. 208–221.

[78] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, "The fact extraction and VERification (FEVER) shared task," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 1–9.

[79] J. Zhou, X. Han, C. Yang, *et al.*, "GEAR: Graph-based evidence aggregating and reasoning for fact verification," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 892–901.

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

[81] H. Wan, H. Chen, J. Du, W. Luo, and R. Ye, "A DQN-based approach to finding precise evidences 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)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1030–1039.

[82] J. Si, D. Zhou, T. Li, X. Shi, and Y. He, "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)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1612–1622.

[83] C. Kruengkrai, J. Yamagishi, and X. Wang, "A multi-level attention model for evidence-based fact checking," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, Aug. 2021, pp. 2447–2460.

[84] Z. Ma, J. Li, G. Li, and Y. Cheng, "GLAF: Global-to-local aggregation and fission network for semantic level fact verification," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 1801–1812.

[85] A. Hanselowski, H. Zhang, Z. Li, *et al.*, "UKP-athene: Multi-sentence textual entailment for claim verification," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 103–108.

[86] K. Jiang, R. Pradeep, and J. Lin, "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)*, 2021, pp. 402–410.

[87] J. Chen, R. Zhang, J. Guo, Y. Fan, and X. Cheng, "GERE: Generative Evidence Retrieval for Fact Verification," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22, New York, NY, USA: Association for Computing Machinery, 2022, pp. 2184–2189.

[88] M. DeHaven and S. Scott, "BEVERS: A general, simple, and performant framework for automatic fact verification," in *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, M. Akhtar, R. Aly, C. Christodoulopoulos, *et al.*, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 58–65.

[89] J. Luken, N. Jiang, and M.-C. de Marneffe, "QED: A fact verification system for the FEVER shared task," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 156–160.

[90] Y. Nie, H. Chen, and M. Bansal, "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*, ser. AAAI'19/IAAI'19/EAAI'19, Honolulu, Hawaii, USA: AAAI Press, 2019.

[91] Y. Nie, S. Wang, and M. Bansal, "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)*, 2019, pp. 2553–2566.

[92] A. Soleimani, C. Monz, and M. Worring, "BERT for evidence retrieval and claim verification," *Advances in Information Retrieval*, vol. 12036, p. 359, 2020.

[93] Z. Liu, C. Xiong, M. Sun, and 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.

[94] W. Zhong, J. Xu, D. Tang, *et al.*, "Reasoning over semantic-level graph for fact checking," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6170–6180.

[95] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[96] G. Bekoulis, C. Papagiannopoulou, and N. Deligiannis, "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*, Online: Association for Computational Linguistics, Jun. 2021, pp. 23–28.

[97] M. Lewis, Y. Liu, N. Goyal, *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[98] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2249–2255.

[99] R. Ghaeini, S. A. Hasan, V. Datla, *et al.*, "DR-BiLSTM: Dependent reading bidirectional LSTM for natural language inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1460–1469.

[100] A. Hanselowski, H. Zhang, Z. Li, *et al.*, "Ukp-athene: Multi-sentence textual entailment for claim verification," *EMNLP 2018*, p. 103, 2018.

[101] Y. Nie, H. Chen, and M. Bansal, "Combining fact extraction and verification with neural semantic matching networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6859–6866.

[102] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced LSTM for natural language inference," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1657–1668.

[103] D. Stammbach and G. Neumann, "Team domlin: Exploiting evidence enhancement for the fever shared task," in *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, 2019, pp. 105–109.

[104] A. Roberts, C. Raffel, K. Lee, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Google, Tech. Rep.*, 2019.

[105] C. Chen, F. Cai, X. Hu, J. Zheng, Y. Ling, and H. Chen, "An entity-graph based reasoning method for fact verification," *Information Processing & Management*, vol. 58, no. 3, p. 102 472, 2021.

[106] C. Chen, F. Cai, X. Hu, W. Chen, and H. Chen, "HHGN: A Hierarchical Reasoning-based Heterogeneous Graph Neural Network for fact verification," *Information Processing & Management*, vol. 58, no. 5, p. 102 659, 2021.

[107] C. Chen, J. Zheng, and H. Chen, "Knowledge-enhanced graph attention network for fact verification," *Mathematics*, vol. 9, no. 16, p. 1949, 2021.

[108] W. Zhong, J. Xu, D. Tang, *et al.*, "Reasoning over semantic-level graph for fact checking," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 6170–6180.

[109] C. Hidey and M. Diab, "Team SWEEPer: Joint sentence extraction and fact checking with pointer networks," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 150–155. DOI: 10.18653/v1/W18-5525. [Online]. Available: https://aclanthology.org/W18-5525.

[110] C. Hidey, T. Chakrabarty, T. Alhindi, *et al.*, "DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 8593–8606.

[111] W. Yin and D. Roth, "TwoWingOS: A two-wing optimization strategy for evidential claim verification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 105–114.

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

[113] S. Subramanian and K. Lee, "Hierarchical Evidence Set Modeling for automated fact extraction and verification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 7798–7809.

[114] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, "HoVer: A dataset for many-hop fact extraction and claim verification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 3441–3460.

[115] A. Sathe, S. Ather, T. M. Le, N. Perry, and J. Park, "Automated fact-checking of claims from Wikipedia," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, *et al.*, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 6874–6882.

[116] W. Chen, H. Wang, J. Chen, *et al.*, "TabFact: A Large-scale Dataset for Table-based Fact Verification," in *International Conference on Learning Representations*, 2019.

[117] V. Gupta, M. Mehta, P. Nokhiz, and V. Srikumar, "INFOTABS: Inference on tables as semi-structured data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 2309–2324.

[118] R. Aly, Z. Guo, M. S. Schlichtkrull, *et al.*, "The fact extraction and verification over unstructured and structured information (feverous) shared task," in *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, 2021, pp. 1–13.

[119] J. Nørregaard and L. Derczynski, "DanFEVER: Claim verification dataset for Danish," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, S. Dobnik and L. Øvrelid, Eds., Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 422–428.

[120] H. Ullrich, J. Drchal, M. Rỳpar, H. Vincourová, and V. Moravec, "CsFEVER and CTKFacts: Acquiring Czech data for fact verification," *Language Resources and Evaluation*, vol. 57, no. 4, pp. 1571–1605, 2023.

[121] S. Lin, W. Hong, D. Wang, and T. Li, "A survey on expert finding techniques," *Journal of Intelligent Information Systems*, vol. 49, no. 2, pp. 255–279, 2017.

[122] M. Z. Al-Taie, S. Kadry, and A. I. Obasa, "Understanding Expert Finding Systems: Domains and Techniques," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–9, 2018.

[123] N. Nikzad–Khasmakhi, M. Balafar, and M. R. Feizi–Derakhshi, "The State-of-the-Art in Expert Recommendation Systems," *Engineering Applications of Artificial Intelligence*, vol. 82, pp. 126–147, 2019.

[124] M. Neshati, Z. Fallahnejad, and H. Beigy, "On Dynamicity of Expert Finding in Community Question Answering," *Information Processing & Management*, vol. 53, no. 5, pp. 1026–1042, 2017.

[125] M. Dehghan, M. Biabani, and A. A. Abin, "Temporal Expert Profiling: With an Application to T-Shaped Expert Finding," *Information Processing & Management*, vol. 56, no. 3, pp. 1067–1079, 2019.

[126] M. Dehghan, H. A. Rahmani, A. A. Abin, and V.-V. Vu, "Mining Shape of Expertise: A Novel Approach Based on Convolutional Neural Network," *Information Processing & Management*, vol. 57, no. 4, p. 102 239, 2020.

[127] S. Yuan, Y. Zhang, J. Tang, W. Hall, and J. B. Cabotà, "Expert Finding in Community Question Answering: A Review," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 843–874, 2020.

[128] N. Nikzad-Khasmakhi, M. Balafar, M. R. Feizi-Derakhshi, and C. Motamed, "Berters: Multimodal representation learning for expert recommendation system with transformers and graph embeddings," *Chaos, Solitons & Fractals*, vol. 151, p. 111 260, 2021.

[129] A. Askari, S. Verberne, and G. Pasi, "Expert Finding in Legal Community Question Answering," in *European Conference on Information Retrieval*, Springer, 2022, pp. 22–30.

[130] Z. Fallahnejad and H. Beigy, "Attention-Based Skill Translation Models for Expert Finding," *Expert Systems with Applications*, vol. 193, p. 116 433, 2022.

[131] D. Wu, S. Fan, and F. Yuan, "Research on Pathways of Expert Finding on Academic Social Networking Sites," *Information Processing & Management*, vol. 58, no. 2, p. 102 475, 2021.

[132] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261–270.

[133] W. Niu, Z. Liu, and J. Caverlee, "Lexl: A learning approach for local expert discovery on twitter," in *European Conference on Information Retrieval*, Springer, 2016, pp. 803–809.

[134] W. Niu, Z. Liu, and J. Caverlee, "On Local Expert Discovery via Geo-Located Crowds, Queries, and Candidates," *ACM Trans. Spatial Algorithms Syst.*, vol. 2, no. 4, Nov. 2016, ISSN: 2374-0353.

[135] W. Li, C. Eickhoff, and A. P. de Vries, "Probabilistic local expert retrieval," in *European Conference on Information Retrieval*, Springer, 2016, pp. 227–239.

[136] C. Liang, Z. Liu, and M. Sun, "Expert finding for microblog misinformation identification," in *Proceedings of COLING 2012: Posters*, 2012, pp. 703–712.

[137] G. Li, M. Dong, F. Yang, *et al.*, "Misinformation-oriented expert finding in social networks," *World Wide Web*, vol. 23, no. 2, pp. 693–714, 2020.

[138] A. Pal and S. Counts, "Identifying Topical Authorities in Microblogs," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 45–54.

[139] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, "Cognos: Crowdsourcing Search for Topic Experts in Microblogs," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '12, Portland, Oregon, USA: Association for Computing Machinery, 2012, pp. 575–590.

[140] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, "Choosing the Right Crowd: Expert Finding in Social Networks," in *Proceedings of the 16th International Conference on Extending Database Technology*, ser. EDBT '13, Genoa, Italy: Association for Computing Machinery, 2013, pp. 637–648.

[141] R. Yeniterzi and J. Callan, "Constructing Effective and Efficient Topic-Specific Authority Networks for Expert Finding in Social Media," in *Proceedings of the first international workshop on Social media retrieval and analysis*, 2014, pp. 45–50.

[142] W. Wei, G. Cong, C. Miao, F. Zhu, and G. Li, "Learning to find topic experts in Twitter via different relations," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1764–1778, 2016.

[143] P. Lahoti, G. De Francisci Morales, and A. Gionis, "Finding Topical Experts in Twitter via Query-Dependent Personalized PageRank," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 155–162.

[144] B. D. Horne, D. Nevo, and S. Adalı, "Recognizing Experts on Social Media: A Heuristics-Based Approach," *SIGMIS Database*, vol. 50, no. 3, pp. 66–84, Jul. 2019, ISSN: 0095-0033.

[145] Z. Cheng, J. Caverlee, H. Barthwal, and V. Bachani, "Who is the barbecue king of texas? a geo-spatial approach to finding local experts on twitter," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 335–344.

[146] D. A. Reynolds, "Gaussian mixture models.," *Encyclopedia of biometrics*, vol. 741, pp. 659–663, 2009.

[147] Y. Ma, Y. Yuan, G. Wang, Y. Wang, D. Ma, and P. Cui, "Local experts finding across multiple social networks," in *International Conference on Database Systems for Advanced Applications*, Springer, 2019, pp. 536–554.

[148] S. James *et al.*, "Finding Experts in Social Media Data using a Hybrid Approach," *arXiv preprint arXiv:2104.03920*, 2021.

[149] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi, "Inferring Who-is-Who in the Twitter Social Network," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 533–538, Sep. 2012, ISSN: 0146-4833.

[150] T. Alshaabi, D. R. Dewhurst, J. R. Minot, *et al.*, "The Growing Echo Chamber of Social Media: Measuring Temporal and Social Contagion Dynamics for over 150 Languages on Twitter for 2009-2020," *CoRR*, vol. abs/2003.03667, 2020.

[151] A. Khalil, M. Jarrah, M. Aldwairi, and Y. Jararweh, "Detecting Arabic Fake News using Machine Learning," in *2021 Second International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, IEEE, 2021, pp. 171–177.

[152] A. Khalil, M. Jarrah, M. Aldwairi, and M. Jaradat, "AFND: Arabic Fake News Dataset for the Detection and Classification of Articles Credibility," *Data in Brief*, vol. 42, p. 108 141, 2022.

[153] K. S. Jones and C. J. Van Rijsbergen, "Information Retrieval Test Collections," *Journal of documentation*, vol. 32, pp. 59–75, 1976.

[154] C. Buckley and E. M. Voorhees, "Evaluating Evaluation Measure Stability," in *ACM SIGIR Forum*, ACM New York, NY, USA, vol. 51, 2017, pp. 235–242.

[155] K. Roberts, T. Alam, S. Bedrick, *et al.*, "Searching for Scientific Evidence in a Pandemic: An Overview of TREC-COVID," *Journal of Biomedical Informatics*, vol. 121, p. 103 865, 2021.

[156] M. Hasanain, R. Suwaileh, T. Elsayed, M. Kutlu, and H. Almerekhi, "EveTAR: Building a large-scale multi-task test collection over Arabic tweets," *Information Retrieval Journal*, vol. 21, no. 4, pp. 307–336, 2018.

[157] M. Hasanain, Y. Barkallah, R. Suwaileh, M. Kutlu, and T. Elsayed, "ArTest: The First Test Collection for Arabic Web Search with Relevance Rationales," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2017–2020.

[158] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[159] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, pp. 159–174, 1977.

[160] K. S. Jones, S. Walker, and S. E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 2," *Information Processing & Management*, vol. 36, no. 6, pp. 809–840, 2000.

[161] S. Wang, S. Zhuang, and G. Zuccon, "BERT-Based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval," in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ser. ICTIR '21, Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 317–324, ISBN: 9781450386111.

[162] H. Li, S. Wang, S. Zhuang, *et al.*, "To Interpolate or Not to Interpolate: PRF, Dense and Sparse Retrievers," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22, Madrid, Spain: Association for Computing Machinery, 2022, pp. 2495–2500.

[163] A. Abolghasemi, A. Askari, and S. Verberne, "On the Interpolation of Contextualized Term-based Ranking with BM25 for Query-by-Example Retrieval," in *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, 2022, pp. 161–170.

[164] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," *arXiv preprint arXiv:1901.04085*, 2019.

[165] R. Nogueira, W. Yang, K. Cho, and J. Lin, "Multi-Stage Document Ranking with BERT," *arXiv preprint arXiv:1910.14424*, 2019.

[166] D. Lin, J. Tang, X. Li, K. Pang, S. Li, and T. Wang, "BERT-SMAP: Paying Attention to Essential Terms in Passage Ranking Beyond BERT," *Information Processing & Management*, vol. 59, no. 2, p. 102 788, 2022.

[167] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, 2020, p. 9.

[168] 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.

[169] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 92–104.

[170] I. A. Abu Farha and W. Magdy, "A Comparative Study of Effective Approaches for Arabic Sentiment Analysis," *Information Processing & Management*, vol. 58, no. 2, p. 102 438, 2021.

[171] A. El Mekki, A. El Mahdaouy, I. Berrada, and A. Khoumsi, "AdaSL: An Unsupervised Domain Adaptation Framework for Arabic Multi-Dialectal Sequence Labeling," *Information Processing & Management*, vol. 59, no. 4, p. 102 964, 2022.

[172] W. Shishah, "JointBert for Detecting Arabic Fake News," *IEEE Access*, vol. 10, pp. 71 951–71 960, 2022.

[173] O. Nael, Y. ELmanyalawy, and N. Sharaf, "AraScore: A Deep Learning-Based System for Arabic Short Answer Scoring," *Array*, vol. 13, p. 100 109, 2022.

[174] W. Mansour, T. Elsayed, and A. Al-Ali, "Did I See It Before? Detecting Previously-Checked Claims over Twitter," in *European Conference on Information Retrieval*, Springer, 2022, pp. 367–381.

[175] I. Sabei, A. Mourad, and G. Zuccon, "SCC - A Test Collection for Search in Chat Conversations," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, ser. CIKM '22, Atlanta, GA, USA: Association for Computing Machinery, 2022, pp. 4429–4433.

[176] E. J. Gerritse, F. Hasibi, and A. P. de Vries, "Entity-Aware Transformers for Entity Search," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1455–1465.

[177] H. Hosseini and E. Bagheri, "Learning to Rank Implicit Entities on Twitter," *Information Processing & Management*, vol. 58, no. 3, p. 102 503, 2021.

[178] R. Reinanda, E. Meij, M. de Rijke, *et al.*, *Knowledge Graphs: An Information Retrieval Perspective*. Now Publishers, 2020.

[179] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web*, Springer, 2007, pp. 722–735.

[180] D. Vrandečić and M. Krötzsch, "Wikidata: A Free Collaborative Knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[181] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[182] S. Tedeschi, S. Conia, F. Cecconi, and R. Navigli, "Named Entity Recognition for Entity Linking: What Works and What's Next," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2584–2596.

[183] I. Yamada, A. Asai, J. Sakuma, *et al.*, "Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 23–30.

[184] F. Hasibi, F. Nikolaev, C. Xiong, *et al.*, "DBpedia-Entity V2: A Test Collection for Entity Search," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '17, Shinjuku, Tokyo, Japan: Association for Computing Machinery, 2017, pp. 1265–1268.

[185] F. Nikolaev and A. Kotov, "Joint Word and Entity Embeddings for Entity Retrieval from a Knowledge Graph," in *Advances in Information Retrieval*, Cham: Springer International Publishing, 2020, pp. 141–155.

[186] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira, "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*, 2021, pp. 2356–2362.

[187] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 2054–2059.

[188] W. Lan, Y. Chen, W. Xu, and A. Ritter, "An Empirical Study of Pre-trained Transformers for Arabic Information Extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 4727–4734.

[189] W. Antoun, F. Baly, and H. Hajj, "AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 191–195.

[190] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in *ICLR*, 2020.

[191] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[192] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[193] O. Obeid, N. Zalmout, S. Khalifa, *et al.*, "CAMeL tools: An open source python toolkit for Arabic natural language processing," in *Proceedings of the 12th language resources and evaluation conference*, 2020, pp. 7022–7032.

[194] M. N. Rateb and S. Alansary, "A Critical Survey on Arabic Named Entity Recognition and Diacritization Systems," in *2022 20th International Conference on Language Engineering (ESOLEC)*, IEEE, vol. 20, 2022, pp. 158–165.

[195] A. Aldumaykhi, S. Otai, and A. Alsudais, "Comparing Open Arabic Named Entity Recognition Tools," *arXiv preprint arXiv:2205.05857*, 2022.

[196] S. Esmeir, "SERAG: Semantic Entity Retrieval from Arabic Knowledge Graphs," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 219–225.

[197] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[198] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[199] C. L. Clarke, G. V. Cormack, and E. A. Tudhope, "Relevance ranking for one to three term queries," *Information processing & management*, vol. 36, no. 2, pp. 291–311, 2000.

[200] C. Macdonald and I. Ounis, "The Influence of the Document Ranking in Expert Search," *Information Processing & Management*, vol. 47, no. 3, pp. 376–390, 2011.

[201] N. Craswell, A. P. De Vries, and I. Soboroff, "Overview of the TREC 2005 Enterprise Track.," in *Trec*, vol. 5, 2005, pp. 1–7.

[202] W. Ferreira and A. Vlachos, "Emergent: A novel data-set for stance classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1163–1168.

[203] T. Alhindi, A. Alabdulkarim, A. Alshehri, M. Abdul-Mageed, and P. Nakov, "AraStance: A Multi-Country and Multi-Domain Dataset of Arabic Stance Detection for Fact Checking," *NLP4IF 2021*, p. 57, 2021.

[204] Z. S. Ali, W. Mansour, T. Elsayed, and A. Al-Ali, "AraFacts: The First Large Arabic Dataset of Naturally Occurring Claims," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 231–236.

[205] Y. Li and C. Scarton, "Revisiting rumour stance classification: Dealing with imbalanced data," in *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 38–44.

[206] J. Zheng, A. Baheti, T. Naous, W. Xu, and A. Ritter, "Stanceosaurus: Classifying Stance Towards Multicultural Misinformation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2132–2151.

[207] J. Khouja, "Stance Prediction and Claim Verification: An Arabic Perspective," in *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Seattle, USA: Association for Computational Linguistics, 2020.

[208] R. Baly, M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, and P. Nakov, "Integrating Stance Detection and Fact Checking in a Unified Corpus," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 21–27.

[209] N. S. Alturayeif, H. A. Luqman, and M. A. K. Ahmed, "Mawqif: A multi-label Arabic dataset for target-specific stance detection," in *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 174–184.

[210] L. H. X. Ng and K. M. Carley, "Is my stance the same as your stance? a cross validation study of stance detection datasets," *Information Processing & Management*, vol. 59, no. 6, p. 103 070, 2022.

[211] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2019, pp. 9260–9269.

[212] A. Baheti, M. Sap, A. Ritter, and M. Riedl, "Just say no: Analyzing the stance of neural dialogue generation in offensive contexts," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4846–4862.

[213] A. Hanselowski, P. Avinesh, B. Schiller, *et al.*, "A retrospective analysis of the fake news challenge stance-detection task," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1859–1874.

[214] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Information Processing & Management*, vol. 60, no. 1, p. 103 120, 2023.

[215] A. Abdelali, H. Mubarak, Y. Samih, S. Hassan, and K. Darwish, "QADI: Arabic dialect identification in the wild," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 1–10.

[216] H. Mubarak and K. Darwish, "Using Twitter to collect a multi-dialectal corpus of arabic," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 1–7.

[217] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, "Cross-domain label-adaptive stance detection," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9011–9028.

[218] I. Abu Farha and W. Magdy, "Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 21–31.

[219] N. Bai, F. Meng, X. Rui, and Z. Wang, "A multi-task attention tree neural net for stance classification and rumor veracity detection," *Applied Intelligence*, vol. 53, no. 9, pp. 10 715–10 725, 2023.

[220] J. Ma, W. Gao, and 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)*, I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1980–1989.

[221] S. Hassan, H. Mubarak, A. Abdelali, and K. Darwish, "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*, 2021, pp. 113–118.

[222] P. Nakov, D. Corney, M. Hasanain, *et al.*, "Automated fact-checking for assisting human fact-checkers," in *IJCAI*, International Joint Conferences on Artificial Intelligence, 2021, pp. 4551–4558.

[223] E. Park, J.-H. Lee, D. Jeon, S. Kim, I. Kang, and S.-H. Na, "SISER: Semantic-infused selective graph reasoning for fact verification," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 1367–1378.

[224] G. Izacard, M. Caron, L. Hosseini, *et al.*, "Unsupervised dense information retrieval with contrastive learning," *arXiv preprint arXiv:2112.09118*, 2021.

[225] C. Zhao, C. Xiong, C. Rosset, X. Song, P. Bennett, and S. Tiwary, "Transformer-xh: Multi-evidence reasoning with extra hop attention," in *International Conference on Learning Representations*, 2019.

[226] J. Chen, Q. Bao, C. Sun, *et al.*, "Loren: Logic-regularized reasoning for interpretable fact verification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 10 482–10 491.

[227] G. Bekoulis, C. Papagiannopoulou, and N. Deligiannis, "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*, A. Feldman, G. Da San Martino, C. Leberknight, and P. Nakov, Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 23–28.

[228] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[229] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang, "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*, ser. SIGIR '23, New York, NY, USA: Association for Computing Machinery, 2023, pp. 2733–2743, ISBN: 9781450394086.

[230] J. Thorne, A. Vlachos, C. Christodoulopoulos, and 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)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 809–819.

[231] S. Ubani, S. O. Polat, and R. Nielsen, "ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT," *arXiv preprint arXiv:2304.14334*, 2023.

[232] L. Pan, Y. Zhang, and M.-Y. Kan, "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)*, 2023, pp. 511–524.

[233] Z. Yue, H. Zeng, Y. Zhang, L. Shang, and D. Wang, "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)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5223–5239.