

# ES-VRAI at CheckThat! 2023: Leveraging Bio and Lists Information for Enhanced Rumor Verification in Twitter

Hamza Tarik Sadouk<sup>1,\*</sup>, Faouzi Sebbak<sup>1</sup> and Hussem Eddine Zekiri<sup>1</sup>

<sup>1</sup>*Ecole Militaire Polytechnique, PO Box 17, 16111 Bordj El Bahri, Algiers, Algeria*

## Abstract

This paper presents our participation in the CLEF2023 CheckThat! Lab [1], focusing on Task 5, which addresses Authority Finding in Twitter [2]. This study addresses the gap in rumor verification by exploring the use of Bio and Lists information from trusted authorities' bios as evidence sources in social media. A Twitter Bio is a brief description or introduction about oneself or an organization displayed on a user's profile, providing a snapshot of their identity, interests, or purpose. Twitter Lists are curated groups of Twitter accounts created by users to organize and categorize specific individuals or topics for easy viewing and engagement. Previous research has primarily focused on propagation networks and user-generated content. The findings highlight the potential of incorporating Bio and Lists information to enhance existing rumor verification systems. Additionally, the study evaluates different retrieval approaches, showing that combining Bio and Lists information leads to effective lexical retrieval. A hybrid approach combining lexical and semantic retrieval further improves performance. These findings contribute to advancing rumor verification methods in social media.

## Keywords

Rumor verification, Authority finding, Lexical retrieval, Semantic retrieval, Hybrid approach.

## 1. Introduction

Previous research on rumor verification in social media has primarily focused on using propagation networks as a source of evidence, examining reply stances [3], reply structures [4], and retweeters profile features [5]. However, according to [6], where they constructed a dataset for the purpose, no prior studies have explored the use of evidence from the timelines of trusted authorities, defined as "entities with the actual knowledge or power to confirm or refute a particular rumor, for rumor verification in social media" [7]. Identifying the stance of relevant authorities regarding rumors can significantly enhance the evidence sources used by current rumor verification systems. Additionally, it can be a useful tool for fact-checkers to automate the process of examining authority tweets for rumor verification. It's important to note that while the stance of authorities can be a valuable source of evidence, it should complement other sources and may not always be entirely reliable for determining the truthfulness of rumors on its own.

---


*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*

\*Corresponding author.

✉ sadouk.hamza.tarik@gmail.com (H. T. Sadouk); faouzi.sebbak@gmail.com (F. Sebbak); houssemzekiri@gmail.com (H. E. Zekiri)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

The task of expert finding, which aims to identify individuals with relevant expertise in a particular domain, has gained significant attention in recent years [8].

Previous research in this area has often focused on incorporating spatial information to address the expert-finding problem. For example, McDonald et al.[9] conducted a study in a medium-sized software company to develop a system that cataloged people’s expertise, aiming to facilitate the process of locating experts. In a similar vein, Dom et al.[10] proposed a graph-based ranking algorithm for expert finding by analyzing email communications. Their approach incorporated link analysis techniques, assigning each email owner a node with a score derived from PageRank and HITS algorithms. Campbell et al.[11] compared two methods for identifying expertise within email communities. The first method focused solely on email text content, while the second method utilized a graph-based ranking algorithm considering both text and communication patterns. The expert search task was introduced in TREC 2005 Enterprise track[12], with the top-performing approach, THUENT0505, utilizing all W3C web part information, email Lists, and in-link anchor text of related files, then restructuring text content to form description files for each potential expert.

Pal and Counts [13] proposed a probabilistic clustering framework that utilized nodal and topical features to identify authoritative experts in a given topic. Ghosh et al.[14] developed the Cognos system, which utilized Twitter Lists to determine user expertise and demonstrated comparable performance to Twitter’s official system in identifying top users for specific topics. Additionally, Wei et al.[15] investigated the use of multiple types of relationships on Twitter to identify experts associated with a particular topic. Studies, such as the one referenced in [16], integrate user-related Twitter data to improve expert finding outcomes. They employ a semi-supervised graph ranking technique to rank expertise levels.

The subsequent sections of this paper are structured as follows: Section 2 provides an overview of the data employed in this study, Section 3 elucidates the ranking metrics to enhance the understanding of the results, Section 4 presents the proposed methodology, Section 5 offers the experimental results and in-depth discussions, and finally, Section 6 concludes the study.

## 2. Data overview

In this section, we present the Authority Finder in Twitter dataset provided as part of the CheckThat! 2023 competition [7, 2].

Given a tweet stating a rumor, we retrieve a ranked list of authorities that may help verify the rumor. The dataset is available in Arabic only, and it’s composed of the following:

- Rumors: The data is in a JSON format. It contains JSON objects representing rumors. Figure 1 illustrates the different entries for each rumor.
- Relevance Judgments (Qrels): The file is a TAB-separated file in TREC format. Each rumor ID is associated with user IDs of the authorities. The file is in the following format:
  - rumor\_id: Unique ID for the given rumor
  - 0: Literally 0 (this column is needed to comply with the TREC format).
  - user\_id: Unique ID for the given authority.

```

{
  rumor_id [unique ID for the rumor]
  tweet_id [unique tweet ID as provided by Twitter]
  tweet_text [tweet text as collected from Twitter]
  category [category of the rumor which is either politics, sports, or health]
}

```

**Figure 1:** Entries of rumor

**Table 1**  
Authority finding dataset summary

Data	Count
Rumors	150
Authorities	1,044
Users	395,231
Twitter Lists	1,192,284

- relevance: 2 if the authority is highly relevant to the rumor (has higher priority to be contacted); 1 if it is relevant; 0 is assumed for all pairs not appearing in the qrels file.
- Users Metadata: It has a collection of 1000 JSON files. Figure 2 represents the format of each file.

```

{
  user_id [the unique user ID as provided by Twitter]
  name [the name of the user, as they've defined it on their profile]
  description [the text of this user's profile description (also known as bio), if the user provided one]
  translated_name [the name of the user translated by us into Arabic]
  translated_desc [the user's profile description translated by us into Arabic]
  following_count [the number of Twitter users this user is following]
  followers_count [the number of Twitter users following this user]
  verified [indicates if this user is a verified Twitter User (1 or 0)]
  lists_count [the number of Twitter lists this user is member of]
  lists_ids [list of unique IDs of the Twitter lists the user is member of if exists]
  collected_Arabic_tweets_ids [unique IDs of Arabic tweets posted by this user (recent at the collection time)]
}

```

**Figure 2:** Entries of user metadata

- Twitter Lists Metadata: a collection of 1000 JSON files with Twitter Lists information (see Figure 3 for the format)

Table 1 provides a general summary of the different files related to this task.

```

{
  list_id [the unique list ID as provided by Twitter]
  name [the name of the list, as defined when creating the list]
  description [a brief description to let users know about the list, if the user provided one]
  translated_name [the name of the list translated by us into Arabic]
  translated_desc [the list's description translated by us into Arabic]
  member_count [the number of users that are part of this list (added as members by the owner)]
  follower_count [the number of users following this list]
  created_at [the UTC datetime that the list was created on Twitter]
  owner_id [unique ID of the user who owns (created) this list]
  owner_name [the name of the user who owns (created) this list, as they've defined it on their profile]
}

```

**Figure 3:** Entries of Twitter list

### 3. Ranking metrics

P@1[17], P@5[18] and nDCG@5[19] are evaluation metrics commonly used in information retrieval and recommendation systems to measure the quality and relevance of ranked Lists. The official evaluation measure is P@5 to evaluate how systems are able to retrieve authorities at the top of a short retrieved list.

- **Precision at K (P@K):** Precision at K measures the proportion of relevant items among the top K items in a ranked list.

$$P@K = \frac{TP}{K} \quad (1)$$

Where:

- TP (True Positives): The number of relevant items in the top K items.
- K: The value of K for which you want to calculate precision.
- **Normalized Discounted Cumulative Gain at K (nDCG@K):** nDCG is a measure that considers both the relevance and the position of items in a ranked list. It gives higher scores to relevant items that appear at the top of the list.

$$nDCG@K = \frac{DCG}{IDCG} \quad (2)$$

Where:

- DCG (Discounted Cumulative Gain): The sum of the relevance scores of the top K items, with decreasing weights, calculated as follows.

$$DCG = \sum_{i=1}^K \frac{rel_i}{\log_2(i+1)} \quad (3)$$

- IDCG (Ideal Discounted Cumulative Gain): The maximum achievable DCG for the given list.
- K: The value of K for which you want to calculate nDCG.

## 4. Proposed method

Our methodology consists of several steps inspired by [20], including preprocessing, creating user documents, indexing and retrieval, and incorporating lexical and semantic scoring for ranking users. Below is a description of each step in the process:

### 4.1. Indexing and Retrieval

We used Pyserini<sup>1</sup>, a Python toolkit for information retrieval, to index our user collection for lexical retrieval. We experimented using different values of BM25 [21] parameters. We set the language to Arabic and performed normalization on both documents and queries. The Lucene Arabic analyzer was employed for stemming and removing stop words. As our queries (rumors) are expressed in tweets, we preprocessed the data by discarding URLs, emojis, and non-Arabic characters. We considered three types of user representation indexes:

- Bio\_index: each user is represented by their translated Twitter profile name and description.
- Lists\_index: each user is represented by concatenating their translated Twitter list names and descriptions.
- Bio\_lists\_index: each user is represented by concatenating their translated Twitter profile name and description with their translated Twitter Lists.

### 4.2. Lexical Retrieval

We employed the BM25 [21] retrieval model to compute the lexical score  $S_x$  for a candidate user  $u$  given a rumor  $r$ , considering only the textual representation of users. The BM25 algorithm calculates a term-weighted score for a given document-query pair as follows:

$$BM25(t_u, t_r) = \sum_{i \in t_r \cap t_u} IDF(q_i) \cdot \frac{(k_1 + 1) \cdot f(q_i, t_u)}{k_1 \cdot ((1 - b) + b \cdot \frac{|t_u|}{avgdl}) + f(q_i, t_u)} \quad (4)$$

where  $t_u$  and  $t_r$  are the textual representations of the user  $u$  and rumor  $r$ , respectively,  $f(q_i, t_u)$  is the frequency of term  $q_i$  in the user document  $t_u$ ,  $|t_u|$  is the length of the user document,  $avgdl$  is the average document length in the collection, and  $IDF(q_i)$  is the inverse document frequency of term  $q_i$ . The parameters  $k_1$  and  $b$  control the term frequency saturation and field-length normalization, respectively.

### 4.3. Initial Retrieval

Assuming authoritative users have more followers and are involved in more Twitter Lists than regular users. We computed the initial score  $S_i(u, r)$  by incorporating the number of Twitter Lists, followers, and users as follows:

$$S_i(u, r) = S_x(u, r) \times \log_2[(l_u + 2) \left(\frac{f_u}{w_u} + 2\right)] \quad (5)$$

---

<sup>1</sup><https://github.com/castorini/pyserini>

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. We added 2 when computing the logarithm of both factors for smoothing in case  $l_u$  or  $f_u$  is zero. In any of those cases, the initial score will fall back to the lexical score.

#### 4.4. Semantic Reranking

Contextualized transformer-based models such as BERT that are pre-trained on large corpora have shown superiority in document reranking. Given that, we also employed a semantic reranking approach based on BERT to re-rank the initial retrieved users, using the same technique as [20] for the generation of input data. Practically we used three pretrained Arabic models:

- **AraBERT** [22], a BERT-based language model, is optimized for Arabic NLP tasks, demonstrating effectiveness in sentiment analysis, and text classification.
- **ArBERT** [23] is a large-scale pre-trained masked language model for Modern Standard Arabic, based on BERT-base architecture, with 163 million parameters.
- **MARBERT** [23] is an Arabic language model pre-trained on a diverse Twitter dataset specifically designed to handle Arabic dialect variations.

#### 4.5. Hybrid Reranking

As the initial score incorporates user profile features like the count of Twitter Lists, followers, and followees, which are crucial in measuring user popularity, we adopted an existing hybrid approach [20], that combines both the initial and semantic scores to compute a final score of candidate users as follows:

$$S_h(u, r) = \alpha \times \hat{S}_i(u, r) + (1 - \alpha) \times S_s(u, r), \quad (6)$$

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. This hybrid approach allows us to leverage both lexical and semantic information along with user profile features to better identify authoritative users for Arabic rumors. Figure 4 shows the whole process.

## 5. Results and discussion

The results presented in this study were obtained using the Development Dataset prior to the release of the Test dataset, with the objective of identifying the most appropriate configuration to be employed during the evaluation phase. This pre-release assessment aimed to discern the optimal setup for subsequent evaluations and ensure the reliability and validity of the findings.

The following tables (2 3 4 5) show the results of the different previously mentioned steps of ranking.

Table 2 (Lexical Retrieval): The table compares the performance of three different index approaches (Bio Index, Lists Index, and Bio Lists Index) using various combinations of the parameters  $k_1$  and  $b$ . We used Greedy-Search to evaluate the optimal configuration of BM25,

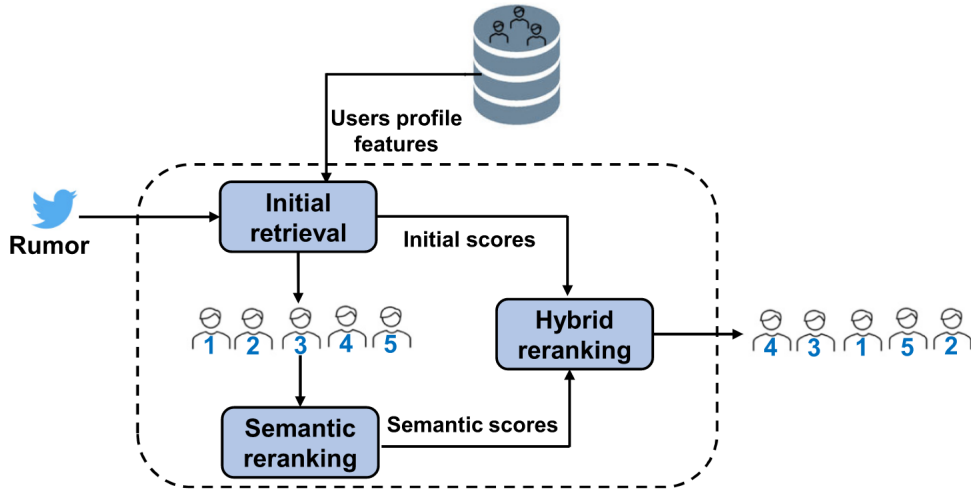


Figure 4: Used authority finding model [20]

Table 2  
Lexical Retrieval

Index	k1	b	P@1	P@5	nDCG@5
<b>Bio Index</b>	1	0.1	0.0667	0.0417	0.0525
<b>Lists Index</b>	1	0.1	0.2500	0.1167	0.1611
	0.9	0.1	0.2250	0.1167	0.1531
	0.9	0.2	0.1917	0.1017	0.1406
	2.9	0.1	0.2500	0.1317	0.1753
	2.9	0.2	0.2417	0.1233	0.1703
	<b>3</b>	<b>0.1</b>	<b>0.2500</b>	<b>0.1350</b>	<b>0.1777</b>
<b>Bio Lists Index</b>	1	0.1	0.2417	0.1167	0.1624
	0.9	0.1	0.2333	0.1117	0.1592
	0.9	0.2	0.1667	0.0900	0.1229
	2.9	0.1	0.2750	0.1383	0.1851
	2.9	0.2	0.2417	0.1100	0.1582
	<b>3</b>	<b>0.1</b>	<b>0.2750</b>	<b>0.1383</b>	<b>0.1865</b>

where  $k1 \in [1, 4]$  with a step of 0.1, and  $b \in [0.1, 0.4]$  with a step of 0.1. The best-performing lexical model was the one utilizing the Bio Lists index and setting the BM25 parameters  $k1$  and  $b$  to 3 and 0.1, respectively, resulting in  $P@1=0.2750$ ,  $P@5=0.1383$ , and  $nDCG@5=0.1865$ . This indicates that combining the Bio and Lists information led to the most effective retrieval strategy. It is important to note that the results for parameter values of  $b = 0.3$  and  $b = 0.4$ , including the default parameters for Pyserini with  $k1 = 0.9$  and  $b = 0.4$ , were not reported in this study due to their relatively moderate performance.

In Table 3, we present the results of the initial retrieval using the Lists and Bio Lists indexes, where we adopted the optimal BM25 parameter settings ( $k1=3$  and  $b=0.1$ ). The Bio Lists Index yielded better results, with  $P@1=0.3500$ ,  $P@5=0.1667$ , and  $nDCG@5=0.2345$ , further

**Table 3**  
Initial Retrieval

Index	k1	b	P@1	P@5	nDCG@5
<b>Lists Index</b>	3	0.1	0.3167	0.1645	0.2223
<b>Bio Lists Index</b>	3	0.1	0.3500	0.1667	0.2345

**Table 4**  
Semantic Retrieval

Model	P@1	P@5	nDCG@5
Initial (Bio+Lists)	0.3500	0.1667	0.2345
AraBERT	0.147	0.133	0.115
ArBERT	0.120	0.113	0.114
MARBERT	0.160	0.107	0.113

**Table 5**  
Hybrid retrieval

Model	P@1	P@5	nDCG@5
Initial (Bio+Lists)	0.3500	0.1667	0.2345
AraBERT	0.3067	0.1802	0.2319
ArBERT	0.3445	0.1745	0.2338
MARBERT	<b>0.3445</b>	<b>0.1835</b>	<b>0.2427</b>

demonstrating the benefit of combining Bio and Lists information and it shows the importance of using user network features.

In Table 4, we present the semantic reranking results using three Arabic-language BERT models (AraBERT, ArBERT, and MARBERT) with the initial retrieval results from Table 2. The initial retrieval outperformed all the semantic models in terms of P@1, P@5, and nDCG@5, suggesting that the lexical approach is more effective than the semantic approach in this case.

The hybrid retrieval results of reranking the top 100 candidate users from the initial retrieval (Bio+Lists) by interpolating initial and semantic scores are presented in Table 5. MARBERT performed the best among the semantic models, with P@1=0.3445, P@5=0.1835, and nDCG@5=0.2427, demonstrating that the hybrid approach combining lexical and semantic retrieval can lead to better performance than using either method alone.

During the evaluation cycle, an attempt was made to employ Hybrid reranking with MARBERT. However, the limited availability of computational resources hindered the production of the necessary results within the specified evaluation cycle deadline. As a result, the submission solely consisted of the results obtained using Initial Retrieval with the optimal configuration of BM25.

## 6. Conclusion

This study evaluates different retrieval approaches for authority finding, where each user is represented by their translated Twitter profile name and description (Bio index) or by the



translated Twitter list names and descriptions (Lists index) or combined (Bio-Lists index). The findings indicate combining Bio and Lists information leads to the most effective lexical retrieval strategy. The initial retrieval process demonstrates the superiority of the Bio Lists Index over the Lists Index, emphasizing the importance of incorporating Bio and Lists information. Moreover, the results showed that both lexical and initial retrieval models outperformed all the semantic retrieval models. However, a hybrid approach combining lexical and semantic retrieval, particularly using the MARBERT model, shows improved performance. These results highlight the benefits of combining different retrieval methods to achieve optimal information retrieval outcomes.

## References

- [1] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, , T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. 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, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, 2023.
- [2] F. Haouari, Z. Sheikh Ali, T. Elsayed, Overview of the CLEF-2023 CheckThat! lab task 5 on authority finding in twitter, in: *Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF '2023*, Thessaloniki, Greece, 2023.
- [3] S. Kumar, K. M. 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*, 2019, pp. 5047–5058.
- [4] J. Choi, T. Ko, Y. Choi, H. Byun, C.-k. Kim, Dynamic graph convolutional networks with attention mechanism for rumor detection on social media, *Plos one* 16 (2021) e0256039.
- [5] Y. Liu, Y.-F. Wu, Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, in: *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [6] F. Haouari, T. Elsayed, Detecting stance of authorities towards rumors in arabic tweets: a preliminary study, in: *European Conference on Information Retrieval*, Springer, 2023, pp. 430–438.
- [7] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, et al., Clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, Springer, 2023, pp. 506–517.
- [8] E. H. Chi, Who knows? searching for expertise on the social web: Technical perspective, *Communications of the ACM* 55 (2012) 110–110.
- [9] D. W. McDonald, M. S. Ackerman, Just talk to me: a field study of expertise location, in:

- Proceedings of the 1998 ACM conference on Computer supported cooperative work, 1998, pp. 315–324.
- [10] B. Dom, I. Eiron, A. Cozzi, Y. Zhang, Graph-based ranking algorithms for e-mail expertise analysis, in: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, 2003, pp. 42–48.
  - [11] C. S. Campbell, P. P. Maglio, A. Cozzi, B. Dom, Expertise identification using email communications, in: Proceedings of the twelfth international conference on Information and knowledge management, 2003, pp. 528–531.
  - [12] N. Craswell, A. P. De Vries, I. Soboroff, Overview of the trec 2005 enterprise track., in: Trec, volume 5, 2005, pp. 1–7.
  - [13] A. Pal, 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.
  - [14] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, 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, 2012, pp. 575–590.
  - [15] W. Wei, G. Cong, C. Miao, F. Zhu, G. Li, Learning to find topic experts in twitter via different relations, *IEEE Transactions on Knowledge and Data Engineering* 28 (2016) 1764–1778.
  - [16] Z. Cheng, J. Caverlee, H. Barthwal, V. Bachani, Finding local experts on twitter, in: Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 241–242.
  - [17] E. Yilmaz, J. A. Aslam, Estimating average precision with incomplete and imperfect judgments, in: Proceedings of the 15th ACM international conference on Information and knowledge management, 2006, pp. 102–111.
  - [18] E. M. Voorhees, The trec question answering track, *Natural Language Engineering* 7 (2001) 361–378.
  - [19] K. Järvelin, J. Kekäläinen, Ir evaluation methods for retrieving highly relevant documents, in: *ACM SIGIR Forum*, volume 51, ACM New York, NY, USA, 2017, pp. 243–250.
  - [20] F. Haouari, T. Elsayed, W. Mansour, Who can verify this? finding authorities for rumor verification in twitter, *Information Processing & Management* 60 (2023) 103366.
  - [21] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., Okapi at trec-3, *Nist Special Publication Sp 109* (1995) 109.
  - [22] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, European Language Resource Association, Marseille, France, 2020, pp. 9–15. URL: <https://aclanthology.org/2020.osact-1.2>.
  - [23] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, 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), Association for Computational Linguistics, Online, 2021, pp. 7088–7105. URL: <https://aclanthology.org/2021.acl-long.551>. doi:10.18653/v1/2021.acl-long.551.