LLM Based Bilingual Rumor Verification Using Evidence From Authorities

Alaaeddin Alia[1], Muhammad Taimoor Khan[2,*]

[1]Heinrich Heine University Düsseldorf, Germany
[2]GESIS - Leibniz Institute for the Social Sciences, Germany

Abstract

The spread of misinformation as rumors is getting more prevalent on social media with its widespread use as access to instant information. Rumors on social media platforms can have damaging consequences unless timely intercepted. The existing studies on rumor verification use linguistic patterns, sentiment orientation and network structures. It require training data preparation and updating the model to stay up to date with newer rumors. However, little attention is paid to benefit from the known trusted and credible authorities to verify rumors. In this study, we address rumor verification on platform X (previously Twitter) by using evidence from the timeline of authority accounts. We propose LLM based bilingual rumor verification for English and Arabic using SBERT and BM25 to retrieve evidence candidates i.e., relevant tweets from the authority timeline, and finetuned XLM-RoBERTa to detect their stance of the rumor. It achieves F1 score of 0.8133 for English and 0.7647 for Arabic to detect stance label for the rumor using evidence candidates. The rumor is verified by weighted aggregation of its stance labels having accuracy of 0.6923 and 0.5769 for Arabic.

Keywords

rumor verification, LLM based rumor verification, rumor evidence stance detection

1. Introduction

In recent years, social media platforms have become the main sources for accessing information, thereby disrupting established outlets such as television and newspapers. Social media platforms provide quick access to unfiltered news and comprise a decentralized opinion landscape that presents multiple perspectives. However, with the increase in access to information, fake news and rumors are also widely spread on social media platforms, including platform X (previously Twitter). Rumor is unverified information that may spread on social media platforms causing misinformation, confusion, and therefore, affecting various areas such as social events, politics, or even personal matters [1]. For example, in 2020, a rumor circulated on Twitter that a popular fast-food chain, has donated to a controversial political campaign, which led to a brief boycott by customers. Although the claim was later debunked, the rumor had already affected the company's public image, demonstrating how quickly a false narrative can demote a brand or individual.

Many studies have been conducted to verify rumors and false news on social media platforms, focusing on the structure of responses, user profiles, linguistic patterns, sentiment orientation and network structure of the rumor [1, 2, 3]. However, little attention has been paid to the role of official authorities in the process of verifying rumors or claims. This is significant given that the authorities are entities that have the knowledge and power to verify rumors as a credible source. They may support or refute a claim through verified evidence [4]. A hybrid model that combines pretrained large language models (LLMs) such as BERT, MARBERT, AraBERT with lexical, semantic and network based features is used to identify authority accounts on Twitter [5]. It motivates the need for a rumor verification

ROMCIR'25: The 5th International Workshop on Reducing Online Misinformation through Credible Information Retrieval, April 10, 2025, Lucca, Italy
*Corresponding author.
Both authors contributed equally.
taimoor.nlp@gmail.com (M. T. Khan)
https://taimoorkhan-nlp.github.io/ (M. T. Khan)
0000-0002-6542-9217 (M. T. Khan)

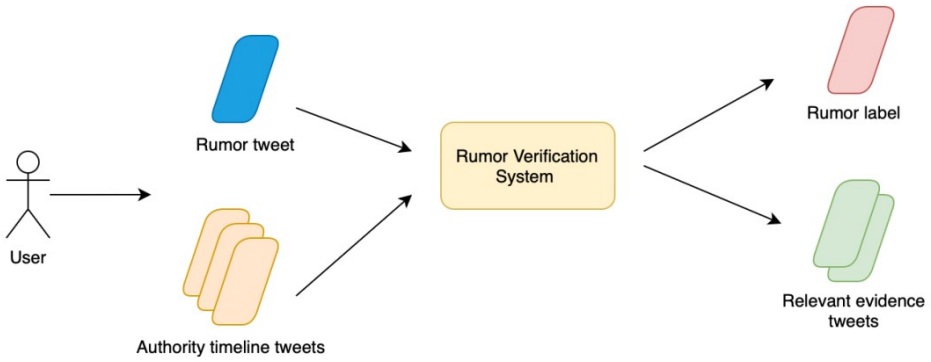


Figure 1: Working of the proposed system as a blackbox that consumes rumor and corresponding authority timeline to decide rumor label and provide evidence tweets for its decision.

system that determine relevant tweets from the authority timelines and use that information to verify rumors, as shown in Figure 1. For example, to verify disease related rumor in a country, their health ministry may be the authority account. Using relevant tweets from this authority timeline can help address the rumor by supporting or refuting it.

A rumor verification system is needed that benefits from the authority timeline tweets as evidence. Each rumor has timelines of the corresponding authority accounts i.e., responsible offices or their representatives as determined in [4, 5]. Although the authority accounts may lack sufficient evidence to confirm or deny a rumor, they are nonetheless assumed to provide correct information. Although, there is evidence of politicians, celebrities and other public figures involved spreading misinformation [6]. The problem statement is that given a rumor and the corresponding authority account(s) timelines, identify relevant tweets for each rumor as evidence candidates, determine stance label of the rumor using each evidence candidate and aggregate to verify the rumor status. A rumor may be supported or refuted based on the available evidence from the authority timeline. In that case, up to 5 evidence tweets are to be provided with the decision that assisted in verifying the rumor status. However, due to lack of conclusive evidence, the rumor is decided as not having enough info.

We propose a bilingual rumor verification system for English and Arabic having four modules. It takes a rumor and the corresponding authority timeline tweets as input and outputs the rumor label and relevant evidence tweets in case the label is *supported* or *refuted*. The first module performs cleaning and preprocessing of all rumors and authority timeline tweets. The second module transforms all rumors and timeline tweets to vectors using dense representation (SBERT) for English and bag-of-words (BoW) sparse representation (BM25) for Arabic. Using cosine similarity, it determines evidence candidates from the authority timelines for each rumor in English and Arabic. The third module uses bilingually finetuned XLM-RoBERTa to detect stance for each rumor and evidence candidate pair. The stances for each rumor can be a mix of supported, refuted or ”not enough info”, depending on the evidence candidates identified in module 2. It performs stance detection for both English and Arabic. Finally, the fourth module performs weighted aggregation of the stance labels to verify the rumor. Our contributions is to devise a large language model (LLM) based pipeline to automatically verify bilingual rumors through reliable authority timeline tweets. In retrieving evidence candidates, SBERT achieved 0.6362 for English and BM25 achieved 0.7833 for Arabic. Finetuned XLM-RoBERTa in stance label detection has the highest F1-score of 0.8133 for English and 0.7647 for Arabic, respectively. Weighted stance label aggregation resulted in rumor verification accuracy of 0.6923 and 0.5769 for English and Arabic, respectively.

2. Literature

Rumor verification: Rumor verification is the process of confirming the veracity of a rumor by gathering evidence, analyzing relevant information, and determining its truthfulness. Various datasets

are available for rumor verification, such as the AuSTR dataset [4], which focuses on the stance of authoritative accounts in Arabic tweets. Another widely used dataset is the FEVER dataset which is designed for fact-checking claims using evidence from Wikipedia [1]. The FEVER dataset shares many similarities with rumor verification tasks in that it challenges systems to classify claims as either supported, refuted, or "not enough info" by retrieving relevant evidence. Both datasets focus on verifying the truthfulness of information using external sources.

Rumor evidence retrieval: Evidence retrieval involves identifying relevant information (evidence documents) from various sources that can either support or refute for a given rumor. Several advanced models have been developed to optimize this process, focusing on retrieving high-quality evidence that improves the accuracy of rumor verification. Kernel Graph Attention Network (KGAT) leverages graph-based structures and kernel-based attention mechanisms to perform fine-grained fact verification, enhancing the model's ability to reason over multiple sources of evidence [2]. This approach constructs an evidence graph in which claims and sentences are nodes, and their relationships are represented as edges. KGAT's ability to capture complex dependencies between pieces of evidence makes it a powerful tool for rumor verification. Evidence-Aware Model focuses on improving sentence retrieval in fact-checking tasks by taking relationships between all potential evidence sentences into account and applies self-attention mechanisms to rank them based on relevance [7]. This evidence-aware approach improves the precision of fact-checking systems by ensuring that only the most relevant sentences are selected for verification.

Text representation: SBERT (sentence bidirectional encoder representations from transformers) is a variation of the original BERT model which is specifically designed to generate meaningful embeddings at the sentence or document level. While BERT produces embeddings for individual tokens (words), SBERT adapts BERT into a Siamese network architecture to compute embeddings that capture the semantic meaning of entire sentences. This is highly effective for tasks such as semantic textual similarity, question answering, and document retrieval [8]. SBERT provides sentence embeddings that can directly be used in downstream tasks such as clustering, ranking, or matching documents based on their meaning. Its advantage over other embedding techniques is its ability to encode the context of a sentence, taking word order and relationships between words into account. TFIDF (term frequency inverse document frequency) is a BoW text vectorization technique that determines word importance through frequency in the document, while penalizing for frequency across most documents. BM25 is also used for information retrieval that improves over TFIDF by using term saturation that restrict frequency, and normalizing for the document length.

Stance detection: Stance detection determines whether the evidence supports, refutes, or provides no clear information about the rumor. [4] introduced the AuSTR as the Arabic rumor tweets dataset and finetuned BERT-based models to classify tweets as agreeing, disagreeing, or unrelated to classify rumors. Coupled hierarchical transformer model perform stance-aware rumor verification in social media conversations [3]. This model captures both local and global interactions within conversation threads and uses a coupled transformer module to integrate stance classification with rumor verification, leading to significant performance improvements. Multi-Level Attention Model for evidence-based fact-checking uses token-level and sentence-level self-attention mechanisms to process and evaluate evidence from multiple sentences [9]. Thereby providing a simple yet effective alternative to more complex graph-based models. XLM-RoBERTa (Cross-lingual language model) is a cross-lingual transformer model built on the RoBERTa architecture trained on 2.5 TB of filtered CommonCrawl data, covering over 100 languages. Through unsupervised learning, XLM-RoBERTa effectively handles a wide range of cross-lingual tasks. While it retains the same architecture as RoBERTa, the fact that it is trained on a more extensive and diverse dataset makes it particularly well-suited for multilingual classification [10]. Knowledge enhanced masked language Model (KE-MLM) is a finetuned BERT-based model aimed at improving stance detection on social media, particularly on Twitter. Instead of random token masking, KE-MLM focuses on stance-relevant tokens identified using the log-odds ratio, thereby improving the model's attention to key contextual words [11].

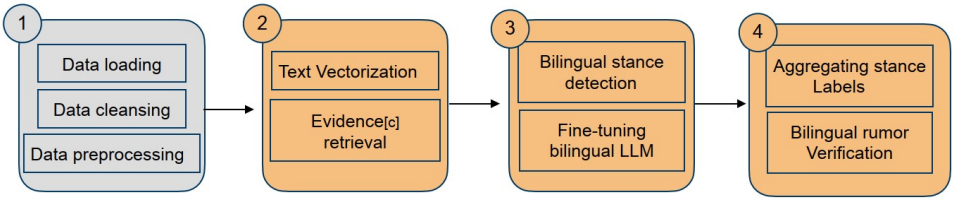


Figure 2: The proposed pipeline consisting of four modules to verify rumor.

3. Methodology

Architecture of the proposed methodology consists of four modules is outlined in Figure 2. The following subsections explain the working of each module.

3.1. Data Preparation

The first module performs data loading, cleansing, and preprocessing. Bad rumors have error codes instead of tweet content in their corresponding timelines, and are removed. This appears to be the data collection problem from the API used. In preprocessing, the tweets are cleaned by removing unwanted characters, hashtags, URLs, mentions etc. It prepares the data for the next module. We also extracted keywords, hashtags, URLs, and emoji embeddings from the rumor and timeline tweets to use as additional features. These features were incorporated to improve the performance of the finetuned stance detection models, thereby allowing us to assess their impact on the overall results.

3.2. Evidence Candidates Retrieval

In this module, the rumors and their corresponding timeline tweets are transformed into their dense embeddings using SBERT model. We also used sparse representation through TFIDF and BM25 with unigrams and bigrams while keeping only the top 1000 most relevant features. The SBERT model has better semantic representation of the data in the embedding vectors that lead to efficient evidence candidates retrieval. Following the text vectorization, we compute the cosine similarity between the SBERT embeddings of the rumors and their respective timeline tweets. Cosine similarity between a rumor and authority timeline tweet can be given as;

$$cos\_sim(rumor_i, tweet_{j,k}) = \frac{rumor_i \cdot tweet_{j,k}}{||rumor_i||\ ||tweet_{j,k}||} \tag{1}$$

Where $rumor_i$ is the i^{th} rumor while $tweet_{j,k}$ is the k^{th} tweet of the j^{th} authority account timeline, corresponding to the i^{th} rumor. It measures the degree of similarity between two vectors, where -1 is complete dissimilarity while 1 is complete similarity. The authority timeline tweets are ordered in decreasing order of their *cos_sim* score with the corresponding rumor. Using fixed threshold as top@k with *k=5,10,15*, evidence candidates i.e., evidence[c] are identified.

3.3. Stance Detection

This module performs bilingual (English and Arabic) multi-label rumor classification using the corresponding evidence candidates. For stance detection, we employ the XLM-RoBERTa transformer-based multilingual model. It is finetuned for the given task using a mix of both English and Arabic samples. A training instance consists of concatenated vectors of rumor and an evidence to predict rumor label. This way, a rumor is paired with all its evidence tweets for a label to increase the training data for better finetuning. We also finetuned KE-MLM model using the same training samples. Due to imbalance in data, we used stratified batches for finetuning these models. This method is especially useful for nuanced decision-making, particularly when certain stance categories, such as the supported, are less

frequent but important, compared to the more common label i.e., ”not enough info”. Other models including both traditional and large language models are used for comparison.

3.4. Rumor Verification

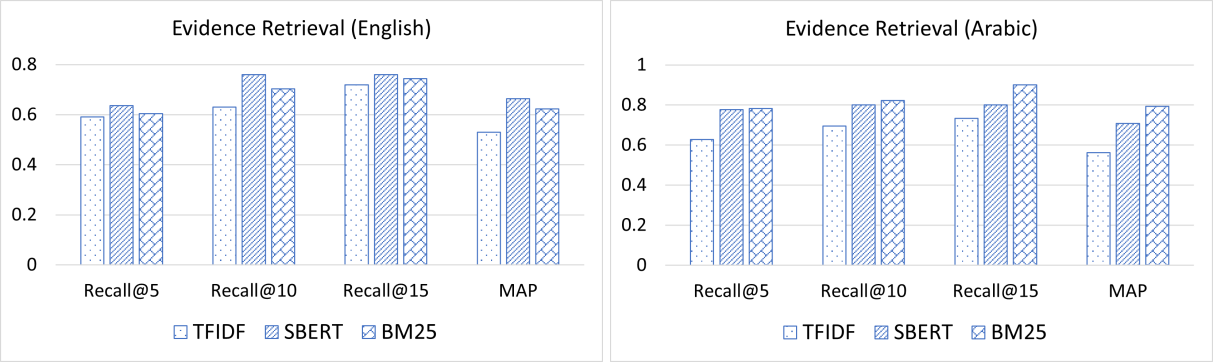
In this module, we aggregate the stance labels produced by the previous module for all pairs of rumors with their corresponding evidence candidates from the authority timelines. Due to the imbalance in data, we use weighted voting aggregation to determine a rumor label from the stance labels of all rumor and evidence candidate pairs. The results are also compared with majority and soft voting aggregation schemes. The weighted scheme assigns weights inversely proportional to the number of instances of a label in the training data. This module verifies the status of the rumor as the final decision. The evidence candidates that helped in determining the label of a rumor as supported or refuted are provided as evidence for it. However, no evidence is needed for the ”not enough info” label.

4. Results

We first split the data into 80% for training and 20% for testing using stratified sampling. This ensured that the ratio of labels remained balanced across both training and test sets. The rumors in the study are independent of one another while the authority timeline tweets in general show higher relevance to the rumor, for mostly covering similar topics. During data cleaning 9 rumors and their corresponding 4,319 timeline tweets are removed from the data for not having meaningful text. These tweets have error codes instead of content that may be sourced from the collection API. The cleaned data has 128 rumors in Arabic that has 53 instances of ”not enough info”, 51 instances of refuted and only 24 instances of supported labels. While, the English data effected by data cleansing has 44 stances of ”not enough info”, 51 of refuted and 24 of supported labels. Both datasets are heavily skewed in favor of ”not enough info” and refuted labels. Since the training data is not enough to finetune XLM-RoBERTa, we separated each training sample into multiple instances by pairing the rumor with all its evidences sharing the same label as provided in the training data. The pairs for each rumor depends on its evidence tweets in the training data (from 1 to 5). No evidence is provided for the rumors labeled ”not enough info” in the training data and therefore, to include them in the finetuning process, randomly sampled tweets from the corresponding authority timeline are used to prepare their training instances. Due to the specialized nature of this approach, the existing rumor datasets i.e., AuSTR and FEVER that do not provide corresponding authority account timeline could not be used for analysis. While, due to bilingual training cost of XLM-RoBERTa, cross validation is expensive and only one time random split is used train the model. The retrieval approaches are evaluated using Recall@k and mean average precision (MAP). Stance classification results are evaluated as F1 score (Micro) and F1 score (Macro).

To evaluate the performance of SBERT model compared to traditional BoW methods in the evidence retrieval task, we use Recall@k with k as 5, 10 and 15 that measures the proportion of relevant evidence to the rumor among the top@k retrieved evidence candidates. MAP assesses the ranking quality by considering the order of the retrieved evidence candidates. The results for both English and Arabic datasets are presented in Figures 3a and 3b. We observe that for the English dataset, the SBERT model achieves the best overall performance, with a Recall@5 of 0.6362, Recall@10 of 0.7607, and Recall@15 of 0.7607. SBERT also outperforms other models in terms of MAP, achieving a score of 0.6635, which indicates that it provides a superior ranking of retrieved evidence. For the Arabic dataset, the BM25 model performs best in terms of Recall across all values of k, reaching Recall@5 of 0.7833, Recall@10 of 0.8222, and Recall@15 of 0.9000. BM25 also achieves the highest MAP score (0.7937), indicating it is particularly effective in ranking relevant evidence in Arabic. While, SBERT performs competitively, with a Recall@5 of 0.7778 and a MAP of 0.7085, demonstrating strong effectiveness across both English and Arabic datasets.

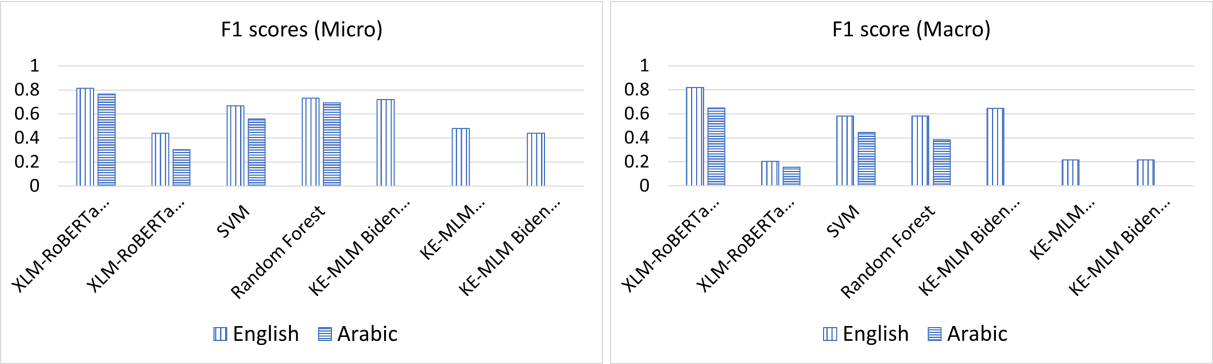
The stance detection performance is evaluated using the F1 Score (Micro) and F1 Score (Macro). We compared the results of our proposed approach with traditional approaches i.e., random forest and SVM and LLM based stance detection models i.e., KE-MLM Trump, KE-MLM Biden. Since these models



(a) English evidence retrieval

(b) Arabic evidence retrieval

Figure 3: Results of evidence retrieval for a rumor from the authority timeline.



(a) F1-score (Micro)

(b) F1-score (Macro)

Figure 4: Results of evidence retrieval for a rumor from the authority timeline.

were finetuned for Trump and Biden tweets respectively, that are different from the our dataset, we also finetuned KE-MLM on the training data called, KE-MLM finetuned. The results of stance detection or classification for English and Arabic are shown in Figure 4a and 4b. Since the present architecture does not represent rumors or their corresponding timeline tweets in a graphical structure, therefore, the results could not be compared with KGAT. The finetuned XLM-RoBERTa model achieves the best performance, with an F1-micro score of 0.8133, and F1-macro score of 0.8179 for English. It suggests that finetuned XLM-RoBERTa model effectively handles stance classes, offering highest accuracy for both English and Arabic. The Random Forest model performs better than SVM, however KE-MLM finetuned model outperform these traditional models. However KE-MLM Trump base, KE-MLM Biden base, and XLM-RoBERTa base did not perform well. This is due to high difference in the nature of our data and the pretraining data of these models. For Arabic, the finetuned XLM-RoBERTa outperforms other models, with F1-micro score of 0.7647, and F1-macro score of 0.6480 Figure 4a and 4b. This highlights the effectiveness of finetuning LLMs for stance detection tasks. SVM and Random Forest also show reasonable performance on the Arabic dataset, with F1 (Micro) scores of 0.5588 and 0.6912, respectively.

The rumor verification is performed through aggregation of the stance labels for each rumor and its corresponding evidence candidate pairs. We compare our weighted aggregation approach addressing imbalance in data with majority voting and soft voting schemes. Weighted voting achieves the highest performance, with F1-micro score of 0.6923 and F1-macro score of 0.6885, shown in Table 1. Majority voting and soft voting both yield the similar F1-micro and F1-macro scores of 0.5769 and 0.5476, respectively. These results indicate that weighted voting is the most effective aggregation scheme for rumor verification with imbalance data. For evidence retrieval, the weighted voting approach outperforms other techniques with a Recall@5 of 0.5556 and a MAP of 0.5556. For Arabic rumor stance

Table 1
Rumor evidence stance labels aggregation using majority, weighted and soft voting schemes.

Aggregation	Data	Rumor Stance Classification		Evidence Retrieval	
		F1 score (Micro)	F1 score (Macro)	Recall@5	MAP
Majority voting	English	0.5769	0.5476	0.3333	0.2889
Weighted voting		0.6923	0.6885	0.5556	0.5556
Soft voting		0.5769	0.5476	0.3333	0.2889
Majority voting	Arabic	0.5000	0.4002	0.3333	0.2889
Weighted voting		0.5769	0.5557	0.6222	0.5778
Soft voting		0.5000	0.4002	0.3333	0.2889

classification weighted scheme achieves F1-micro score of 0.5769, a F1-macro score of 0.5557 with a Recall@5 of 0.6222, and a MAP of 0.5778 for evidence retrieval. Majority voting and soft voting both reach a F1-micro score of 0.5000, F1-macro score of 0.4002, and have lower Recall@5 and MAP values of 0.3333 and 0.2889, respectively.

5. Discussion

The results highlight important differences in model performance across the evidence retrieval, stance detection, and rumor verification tasks for both English and Arabic. For evidence retrieval, the SBERT model demonstrated superior performance on the English, particularly in terms of Recall@5 and MAP. This suggests that SBERT is more effective at capturing semantic similarities for ranking relevant evidence in English, which is likely due to its deep contextualized embeddings. Conversely, for the Arabic dataset, the BM25 model outperformed other models, achieving the highest Recall@15 and MAP scores. This indicates that traditional retrieval techniques such as BM25 are still highly effective for Arabic text, potentially due to the language's morphological richness, which enables simple frequency-based methods to effectively capture relevance. In stance detection, the finetuned XLM-RoBERTa model consistently achieved the best results across both English and Arabic, which suggests that domain-specific finetuning of transformer-based models significantly improves the ability to distinguish between stance classes. However, despite finetuned on equal instances and similar topics for both English and Arabic, the accuracy for English is higher than Arabic. It attributes to either the evidence candidates used for finetuning are not very relevant and/or better English samples were used in pretraining the model. It is interesting to note that traditional models such as SVM and Random Forest, while performing reasonably well, however, were clearly outperformed by XLM-RoBERTa, especially in terms of F1-macro scores. This indicates that XLM-RoBERTa is better at handling class imbalances and providing a more balanced prediction across all classes.

For rumor verification, the aggregation technique of weighted voting proved to be the most effective for both English and Arabic. In particular, weighted voting achieved the highest F1-micro score and F1-macro scores, outperforming both majority voting and soft voting. Due to the imbalance in data, the label weights were inversely proportional to their representation in the training data. The majority and soft voting schemes have similar results for both English and Arabic datasets. In fact, majority voting and soft voting yielded same results, with lower accuracy and F1-macro scores. The majority and soft voting have same score indicating that there is no higher difference in stance intensities of rumor evidence candidate pairs for a corresponding rumor. The results suggest that weighted voting is particularly beneficial in handling cases in which some stance classes are more prevalent, thereby helping to mitigate the impact of class imbalance. There are some limitations to the current approach. The additional feature such as emoji embeddings, hashtags and URLs did not improve the results of stance detection task, which requires more effort for better representation and concatenation with the content embeddings. Moreover, The retrieval mechanism did not consider the presence of stance in the authorities' timeline and therefore did not provide a clear separation between the irrelevant timeline tweets and evidence candidates. Finetuning SBERT for the task may also have improved the evidence

candidates retrieval mechanism. The results show that transformer-based models such as SBERT and XLM-RoBERTa are effective for evidence retrieval and stance, particularly when finetuned for the task. Nevertheless, traditional models such as BM25 remain competitive, particularly for non-English data, and weighted voting emerges as an important technique for improving rumor verification performance.

6. Conclusion

In this research, we addressed rumor verification issue on social media platforms, when there are known authority accounts corresponding to the rumor topic. The proposed system can be deployed as a first-hand rumor detector to alert on rumor tweets with claims that are not supported by the corresponding authority accounts. The proposed methodology centers on utilizing evidence retrieved from authority timelines and stance detection using transformer based pipeline. The results show that SBERT and finetuned XLM-RoBERTa, achieve superior performance for evidence retrieval and stance detection. Our findings emphasize the growing importance of transformer-based models for NLP tasks, while also highlighting areas where traditional methods and aggregation schemes, such as weighted voting, can still play a valuable role. In future, the retrieval module can be improved using evidence-aware model to take relationship among timeline tweets into consideration as well. Feature extraction and their utilization can also be improved to benefit additional features within tweet content. Further exploration of hybrid models that combine traditional retrieval methods, such as BM25, with deep learning techniques could yield promising results, particularly in multilingual or domain-specific contexts. Similarly, enhancing aggregation schemes, such as adaptive weighting schemes based on context, could further boost performance in rumor verification tasks.

References

[1] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and verification (fever) shared task, arXiv preprint arXiv:1811.10971 (2018).

[2] Z. Liu, C. Xiong, M. Sun, Z. Liu, Fine-grained fact verification with kernel graph attention network, arXiv preprint arXiv:1910.09796 (2019).

[3] J. Yu, J. Jiang, L. M. S. Khoo, H. L. Chieu, R. Xia, Coupled hierarchical transformer for stance-aware rumor verification in social media conversations, in: In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020.

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

[5] F. Haouari, T. Elsayed, W. Mansour, Who can verify this? finding authorities for rumor verification in twitter, Information Processing & Management 60 (2023) 103366.

[6] J. S. Brennen, F. M. Simon, P. N. Howard, R. K. Nielsen, Types, sources, and claims of covid-19 misinformation (2020).

[7] G. Bekoulis, C. Papagiannopoulou, 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, 2021, pp. 23–28.

[8] N. Reimers, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[9] C. Kruengkrai, J. Yamagishi, X. Wang, A multi-level attention model for evidence-based fact checking, arXiv preprint arXiv:2106.00950 (2021).

[10] A. Conneau, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[11] K. Kawintiranon, L. Singh, Knowledge enhanced masked language model for stance detection, in: Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies, 2021, pp. 4725–4735.

Appendix

Table 2
Statistical analysis of English and Arabic datasets

Metric	English dataset	Arabic dataset
Total Instances (Rumors)	128	128
Avg. Timelines Tweets per Rumor	216.62	216.62
Min. Tweets in Timeline	1	1
Max. Tweets in Timeline	1,910	1,910
Avg. Evidence per Rumor	2.6	2.6
Avg. Rumor Tweet Length	196.23	150.16
Avg. Timeline Tweet Length	165.4	153.28
Avg. Evidence Tweet Length	261.2	195.0
Percentage of URLs in Rumors	69.75%	70.31%
Percentage of URLs in Timelines	60.36%	80.4%
Percentage of URLs in Evidence Tweets	63.66%	64.86%
Percentage of Hashtags in Rumors	40.62%	41.41%
Percentage of Hashtags in Timelines	47.29%	56.68%
Percentage of Hashtags in Evidence Tweets	49.25%	50.45%

Table 3
English and Arabic total and label-wise affected tweets after each preprocessing step.

English Dataset				
Preprocessing Step	Total	NOT ENOUGH INFO	REFUTES	SUPPORTS
URLs Removed	76.4%	78.2%	75.0%	76.4%
Noise Words Removed	10.0%	5.5%	5.1%	3.1%
Special Characters Removed	99.8%	33.3%	33.3%	33.3%
Emojis Removed	32.4%	8.5%	8.2%	6.2%
Hashtags Removed	59.8%	11.8%	12.4%	10.5%
Mentions Removed	25.1%	4.0%	4.6%	3.1%
Arabic Dataset				
URLs Removed	80.3%	84.8%	75.4%	77.4%
Noise Words Removed	10.1%	5.5%	5.1%	3.1%
Special Characters Removed	99.0%	33.1%	33.0%	32.4%
Emojis Removed	28.4%	6.3%	8.3%	6.3%
Hashtags Removed	56.7%	10.5%	12.6%	10.7%
Mentions Removed	24.2%	3.8%	4.6%	3.1%
Spaces Removed	71.3%	10.6%	9.8%	9.5%

Table 4
Performance of evidence retrieval models on English and Arabic datasets using their respective similarity metrics.

English Dataset					
Representation	Similarity Metric	Recall@5	Recall@10	Recall@15	MAP
TF-IDF	cos sim	0.5911	0.6311	0.7200	0.5296
SBERT	cos sim	**0.6362**	**0.7607**	**0.7607**	**0.6635**
BM25	BM25 sim	0.6044	0.7029	0.7436	0.6226
Arabic Dataset					
TF-IDF	cos sim	0.6278	0.6944	0.7333	0.5623
SBERT	cos sim	0.7778	0.8000	0.8000	0.7085
BM25	BM25 sim	**0.7833**	**0.8222**	**0.9000**	**0.7937**

Table 5
Stance detection performance of different models on the English and Arabic datasets.

	English Dataset		Arabic Dataset	
Model	F1-micro	F1-macro	F1-micro	F1-macro
XLM-RoBERTa Finetuned	**0.8133**	**0.8179**	**0.7647**	**0.6480**
XLM-RoBERTa Base	0.4400	0.2037	0.3033	0.1552
SVM	0.6667	0.5803	0.5588	0.4436
Random Forest	0.7333	0.5816	0.6912	0.3839
KE-MLM Biden Finetuned	0.7200	0.6465		
KE-MLM Trump Base	0.4800	0.2162		
KE-MLM Biden Base	0.4400	0.2157		