Expert Systems With Applications 243 (2024) 122791

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Predicting and analyzing the popularity of false rumors in Weibo

Yida Mu [a], Pu Niu [b,c,*], Kalina Bontcheva [a], Nikolaos Aletras [a]

[a] Department of Computer Science, The University of Sheffield, Sheffield, UK
[b] School of Marxism, Henan University, Kaifeng, China
[c] Research Institute of Marxism, Henan University, Kaifeng, China

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.8374169

Keywords:
Social media
Rumor popularity
Misinformation analysis

ABSTRACT

Malicious online rumors with high popularity, if left undetected, can spread very quickly with damaging societal implications. The development of reliable computational methods for early prediction of the popularity of false rumors is very much needed, as a complement to related work on automated rumor detection and fact-checking. Besides, detecting false rumors with higher popularity in the early stage allows social media platforms to timely deliver fact-checking information to end users. To this end, we (1) propose a new regression task to predict the future popularity of false rumors given both post and user-level information; (2) introduce a new publicly available dataset in Chinese that includes 19,256 false rumor cases from Weibo, the corresponding profile information of the original spreaders and a rumor popularity score as a function of the shares, replies and reports it has received; (3) develop a new open-source domain adapted pre-trained language model, i.e., BERT-Weibo-Rumor and evaluate its performance against several supervised classifiers using post and user-level information. Our best performing model (KG-Fusion) achieves the lowest RMSE score (1.54) and highest Pearson's r (0.636), outperforming competitive baselines by leveraging textual information from both the post and the user profile. Our analysis unveils that popular rumors consist of more conjunctions and punctuation marks, while less popular rumors contain more words related to the social context and personal pronouns. Our dataset is publicly available: https://github.com/YIDAMU/Weibo_Rumor_Popularity.

1. Introduction

Social media platforms (e.g., Twitter, Facebook and Weibo) play an important role in information dissemination related to important events, social emergencies and natural disasters (Castillo, 2016; Imran, Castillo, Diaz, & Vieweg, 2015; Middleton, Middleton, & Modafferi, 2013; Wang, Wang, Ye, Zhu, & Lee, 2016). However, online rumors (i.e., posts with unverified veracity) have been shown to spread faster than reliable information and can thus mislead the public especially when ultimately proven false (Vosoughi, Roy, & Aral, 2018).

The timely publication of fact-checks of such false rumors[1] can both raise user awareness and help prevent rumors from spreading further (Vo & Lee, 2020). Vo and Lee (2019) showed that debunked tweets (i.e., Twitter posts containing false rumors) are more likely to be deleted and for their original spreaders to be suspended. To combat false rumors spreading in social media, independent (e.g., PolitiFact[2]) or in-house (e.g. Weibo) fact-checking platforms have been created with the purpose to debunk suspicious posts.

Fig. 1 shows an example of a debunked false rumor on the Weibo fact-checking platform.[3] The top box shows the rumor: *"Hua Chunying (i.e., the Foreign Ministry Spokesperson of PRC) announces a ban on Chinese stars "*. The exclamation mark sign '!' in the top box denotes that 'This is a debunked rumor' and users can click on it[4] to get relevant fact-checking information. The blue box (middle left) denotes the information of the user reported the rumor and the number of all reports from different users. The orange box (middle right) displays

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author at: School of Marxism, Henan University, Kaifeng, China.
E-mail addresses: y.mu@sheffield.ac.uk (Y. Mu), niupu@henu.edu.cn (P. Niu), k.bontcheva@sheffield.ac.uk (K. Bontcheva), n.aletras@sheffield.ac.uk (N. Aletras).

[1] In this work, we use the term 'false rumor' to refer to any rumors that have been proven false (e.g., unreliable stories on Weibo).
[2] https://www.politifact.com
[3] http://weibo.com/1074273855/IwFJ6dsuQ
[4] https://service.account.weibo.com/show?rid=K1CaS8wtk7K4k

https://doi.org/10.1016/j.eswa.2023.122791
Received 20 April 2023; Received in revised form 27 November 2023; Accepted 27 November 2023
Available online 1 December 2023

the information of the user who posted the false rumor. The green box (bottom) indicates that the original post is debunked as a false rumor and the user (in the orange box) who posted the false rumor will lose 10 'credit points' and cannot post or be followed for the next 15 days.

Fact-checking platforms typically verify such rumors manually, which is highly reliable but expensive in terms of time and costs (Pavleska, Školkay, Zankova, Ribeiro, & Bechmann, 2018; Vo & Lee, 2018). Therefore, fact checkers are increasingly being assisted by automated rumor detection and veracity (i.e., whether a rumor is true or false) prediction systems for retrieving rumor related information more efficiently (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018). To further improve their efficiency, professionals also need a way to prioritize for debunking those detected rumors which are likely to become highly popular and reach a large audience (Parikh, Patil, Makawana, & Atrey, 2019; Smith & Bastian, 2022).

To address the latter challenge, the focus of this paper is on developing computational methods for predicting the popularity of false rumors as soon as they are detected. Identifying false rumors with a higher impact in the early stage allows social media platforms to timely deliver fact-checking information to the public. Previous work has focused on predicting the popularity of social media posts (e.g., tweets, YouTube videos) (Gao et al., 2021; Trzciński & Rokita, 2017) and individual social media users (Lampos, Aletras, Preoţiuc-Pietro, & Cohn, 2014) with applications in advertising and recommend systems. For the purpose of our task, we only consider immediately available contextual information (e.g., rumor content and user profile information), which is crucial for the early detection of false rumors with high popularity. To the best of our knowledge, the regression task of false rumor popularity prediction (i.e., early detection of popularity) has yet to be explored.

To this end, we pose the following three research questions:

- RQ_1 How can we define the popularity of false rumors on Weibo?
- RQ_2 How can we predict the future popularity of false rumors based on post and user-level information?
- RQ_3 What are the most important markers that correlate with high and low-popularity rumors?

To answer these research questions:

- We develop a new publicly available dataset[5] from Weibo, which includes 19,256 debunked false rumors in Chinese associated with a popularity score and Meta-features;
- We evaluate several supervised models using post and user-level information and their combination. Combining these two sources of information with our new pre-trained language model (i.e. BERT-Weibo-Rumor) achieves the best overall performance. Given that most fact-checking platforms (e.g., Weibo rumor debunking platform, PolitiFact, etc.) rely on human resources to manually check the veracity of rumors, social media platforms can first address false rumors with higher popularity.;
- We perform a linguistic analysis to unveil the characteristics of highly popular false rumors compared to those with low popularity. We also unveils that some user profile characteristics (e.g., verified status, number of followers and number of historical posts) are positively correlated with the future popularity of false rumors.

Paper outline The rest of the paper is organized as follows. In Section 2, we discuss previous work related to rumor detection. The task description is introduced in Section 3. We describe the development of our Weibo dataset in Section 4. We discuss the model details, hyperparameters tuning, and results in Section 5. In Section 6, we conduct extensive analysis (including ablation study, error analysis, and qualitative analysis of model prediction to gain insights for future work. We also discuss the ethical considerations, theoretical and practical implications of this work in Section 7. Finally, we conclude and sum up some future directions in Section 8.

[5] Our dataset and source code will be publicly released.

2. Related work

2.1. Rumor detection

A rumor is generally defined as any social media post whose credibility is yet to be verified at the time it was published (Zubiaga et al., 2018). Typically, rumor detection systems first predict if a given post is a rumor or not, and second whether it is true or false. Prior work on automatic rumor detection generally falls into one of the following categories:

- feature-based methods that rely on linguistic (e.g., text) and visual (e.g., images) information to detect unreliable posts (Choi, Oh, Chun, Kwon, & Han, 2022; Karmakharm, Aletras, & Bontcheva, 2019; Qi, Cao, Yang, Guo, & Li, 2019; Rashkin, Choi, Jang, Volkova, & Choi, 2017; Yang, Wang, Wang, & Meng, 2022);
- knowledge-based methods that leverage external knowledge (e.g., Wikipedia) to determine rumor veracity (Hu et al., 2021; Jiang, Liu, Zhao, Sun, & Zhang, 2022; Sun et al., 2023; Wan, Wang, Pang, Wang, & Min, 2023; Wei, Hu, Zhou, Yue, & Hu, 2021);
- user propagation-based methods that consider the diffusion of the rumor (e.g., time-series analysis) (Chen, Zhou, Zhang, & Bonsangue, 2021; Lin et al., 2021; Nobre, Ferreira, & Almeida, 2022).
- early rumor detection methods (Silva, Han, Luo, Karunasekera, & Leckie, 2021; Xia, Xuan, & Yu, 2020; Zhou et al., 2022) that aim to detect rumors as soon as they are posted online. These approaches tend to employ a combination of features (e.g., user features and time-series data) from different periods during the rumor propagation cycle to detect the earliest point in time that a particular post has actually become a rumor (Yuan, Ma, Zhou, Han, & Hu, 2020; Zhou, Shu, Li, & Lau, 2019).

Automatic rumor detection methods are usually evaluated on existing annotated datasets, e.g., Weibo (Ma et al., 2016; Ma, Gao, & Wong, 2017), FEVER (Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018), Twitter15&16 (Ma et al., 2016), PHEME (Zubiaga et al., 2016) and LIAR (Wang, 2017). Kochkina et al. (2023), Mu, Bontcheva, and Aletras (2023), Mu, Song, Bontcheva, and Aletras (2023) evaluate the generalizability of neural-based rumor classifiers across different benchmarks. Recently, interpretable rumor detection methods (e.g., attention-based and rule-based) have also been explored (Atanasova, Simonsen, Lioma, & Augenstein, 2020, 2022; Ayoub, Yang, & Zhou, 2021; Silva et al., 2021), for generating explanations in aid of fact-checking by highlighting evidence. Some of these methods are often embedded in real-world fact-checking platforms e.g., Propagation2Vec (Silva et al., 2021) and Defend (Shu, Cui, Wang, Lee, & Liu, 2019).

Apart from modeling individual posts, previous work has also explored modeling user behavior, e.g. analyzing user reactions and stance towards unreliable posts (Bazmi, Asadpour, & Shakery, 2023; Glenski, Weninger, & Volkova, 2018; Mu & Aletras, 2020; Mu, Niu, & Aletras, 2022) to show that a higher percentage of human users retweet news posts from credible sources (e.g., @BBC and @Reuters) as compared to bots.

2.2. Modeling popularity in social media

Another strand of related work has focused on predicting the popularity of multimodal online content, e.g., YouTube Videos (Kong, Rizoiu, Wu, & Xie, 2018; Pinto, Almeida, & Gonçalves, 2013), tweets (Zhao, Erdogdu, He, Rajaraman, & Leskovec, 2015), Facebook posts (Trzciński & Rokita, 2017) and Weibo posts (Bao, Shen, Huang, & Cheng, 2013; Gao, Ma, & Chen, 2014).

Fig. 1. False rumor debunking pipeline on the Weibo fact-checking platform (English translation also included in the main body of the paper).

Table 1
Specifications of existing Weibo-based rumor datasets. *Note.* ‡ denotes that our dataset contains more user-level features e.g., 'user Credit Score', '# of Likes Received' (i.e., user attributes features (U) from U_6 to U_{12} in Table 2).

Dataset	# of false rumors	Time span	User-level features	# of engagements
Ma et al. (2016)	2,313	2012–2016	✓	✓
Jin, Cao, Guo, Zhang, and Luo (2017)	4,749	2012–2016	x	x
Rao, Miao, Jiang, and Li (2021)	3,034	2016–2021	✓	✓
Song et al. (2021)	1,538	N/A	x	x
Lu, Fan, Song, and Fang (2021)	1,975	2012–2020	x	x
WeiboRumors (Ours)	**19,256**	**2010–2021**	✓ ‡	✓ ‡

Existing work usually relies on post's user engagement metrics (e.g., shares, replies, views, likes, etc.) to represent its popularity (Gao et al., 2021; Yan, Tan, Gao, Tang, & Chen, 2016). Another metric is engagement rate which is calculated as the sum of the user engagement metrics received divided by the number of views of the post (Alkhodair, Fung, Ding, Cheung, & Huang, 2020). To model the popularity score of online posts, post-level features (e.g., textual and visual information) (Pinto et al., 2013; Piotrkowicz, Dimitrova, Otterbacher, & Markert, 2017) and user features (e.g., profile information) (Gelli, Uricchio, Bertini, Del Bimbo, & Chang, 2015; Li, Situ, Gao, Yang, & Liu, 2017; McParlane, Moshfeghi, & Jose, 2014) are commonly used as they are publicly available.

At the level of individual users, Weng, Lim, Jiang, and He (2010) and Lampos et al. (2014) quantify Twitter user impact as a function of the number of followers and friends. They both predict and analyze user impact through user profile and post-level features.

Rumor popularity Previous work on predicting the future popularity of false rumors is limited. Alkhodair et al. (2020) present a classification task for predicting the engagement rate of tweets (high, moderate and low) through solely textual information. However, the predicted engagement rate which is calculated by dividing the sum of the engagement by the sum of the views received on the post, cannot be applied for Weibo posts as it requires the number of views on the post which is not available through the official Weibo API. Similar to Alkhodair et al. (2020), Jiang, Wang, et al. (2022) first categorize rumors into three types based on the degree of user participation (i.e., low, moderate, and high), and then employ user and news interactions to conduct the task of rumor popularity. Parikh et al. (2019) define the impact of online false news articles based on three metrics including (i) the topic of the news items (e.g., politics, economics, science, etc.); (ii) the reputation of the news website that posted the news and (iii) the proliferator's popularity, i.e., the number of followers of users who shared the false news.

2.3. Our work

We note that while some rumors do spread widely (i.e. gain a lot of attention), many others only reach a very small audience. Therefore, it is equally important to detect the **future popularity of false rumors** on social media, so that they can be prioritized for debunking. Given that most fact-checking platforms (e.g., Weibo rumor debunking platform, PolitiFact, etc.) rely on human resources to manually check the veracity of rumors, social media platforms can first address false rumors with higher popularity. Note that we only use information that is immediately available which is crucial for the early detection of false rumors

with high popularity. This task is yet to be explored in computational social science.

3. Task description

We define false rumor popularity prediction as a regression task. Given a false rumor $X = \{(R, P, U)\}$ consisting of textual information R (i.e., a sequence of tokens representing the actual rumor), user profile description P (i.e., a sequence of tokens representing the personal description provided by the user) and user attributes U (e.g., number of followers, posts, etc.), we aim to learn a supervised function f that can predict the popularity score Y of a false rumor. The value of the popularity score is calculated using rumor engagement attributes E, which include the number of shares, number of replies, and number of reports, based on Eq. (1).

4. Data

4.1. Data collection

For our experiments, we create a new dataset using the fact-checking platform provided by Weibo.[6] We opted using Weibo since it is the largest Chinese-based social media platform and its fact-checking platform has enabled the development of many rumor detection datasets (Ma et al., 2016; Rao et al., 2021).

However, these previously published datasets are relatively small (e.g., there are 2,313 and 3,034 false rumor cases from Ma et al. (2016) and Rao et al. (2021) datasets, respectively) and lack the **metadata information** required for our task. For instance, the Song et al. (2021) and Lu et al. (2021) datasets are not suitable for the regression task of predicting the level of popularity of false rumors, as they do not provide information on the number of engagements (e.g., Shares, replies, etc.) received by rumors. We further elaborate on the details of previously publicly available datasets in Table 1.

The Weibo fact-checking platform allows end-users to report suspicious posts (i.e., rumors), which are subsequently checked by professional journalists to verify their veracity and provide fact-checking information. In cases where the information of a post is deemed to be *false*, it is also flagged as a false rumor including information that refutes any claims that it contains. Note that rumors are usually defined as online posts whose veracity is yet to be verified at the time of posting (i.e., they can ultimately turn out to be *true*, *false* or *not verifiable*) (Zubiaga et al., 2018). However, in our dataset all rumors are *false*, i.e., the source post contains debunked false information (see Fig. 1).

We collect a total of 40,936 cases of false rumors using the official Weibo API.[7] All cases have been debunked and cover a period between May 2012 and November 2021.

4.2. Rumor information

For each false rumor, we collect rumor engagement attributes (E), the rumor content (R), the user profile description (P) and user attributes (U).

Rumor engagement attributes (E) include the number of shares (E_1), replies (E_2) that the post received, and the number of times users have reported (E_3) the post on the fact-checking platform. Note that, one rumor can be reported by different users in the Weibo fact-checking platform.

We also collect the text of the false rumor (R) and the user profile description (P). User attributes (U) consist of (1) user social connections (from U_1 to U_5) including the number of followers (i.e., other users who follow this Weibo account), followees (i.e., one can follow other users), bi_followers (i.e., users who follow each other), statuses (i.e., historical posts) and favorites (i.e., one can like posts from other users); (2) user engagement (from U_8 to U_{12}) information including the number of posts, and the number of reactions (e.g., shares, replies, etc from other users) received; (3) Weibo account attributes e.g., Verified Status (U_7) (i.e., Verified or Unverified) and Credit Score (U_6). Note that the 'Credit Score' (U_6) is a unique user-level attribute on Weibo. Weibo users lose some of their credit score for posting false rumors. When a user's credit score falls below a certain threshold, they are not able to post for a period of time.

Table 2
Information associated with each false rumor in our dataset.

Features	Description
Rumor engagement attributes (E)	
$E1$	# of shares
$E2$	# of replies
$E3$	# of reports
Rumor content	
R	Text representing the actual rumor
User profile description	
P	Text describing user's bio
User attributes (U)	
U_1	# of followers (i.e., other users who follow this account)
U_2	# of followees (i.e., one can follow others)
U_3	# of Bi_Followers (i.e., users who follow each other)
U_4	# of statuses (i.e., the number of posts)
U_5	# of favorites (i.e. one can like posts from other users)
U_6	Credit score
U_7	Verified status (i.e., Verified or Unverified)
U_8	# of shares received
U_9	# of likes received
U_{10}	# of replies received
U_{11}	# of likes received in replies
U_{12}	# of all reactions received (i.e., the sum of U_8, U_9, U_{10}, U_{11})

We only consider the rumor content (R), user profile description (P) and user attributes (U) as they are immediately available when false rumors are published on Weibo. These features can be used to train predictive models for detecting highly popular false rumors in an early stage. Table 2 shows a summary of all information collected for each rumor.

4.3. Defining false rumor popularity on Weibo

In social networks, the user engagement (e.g., shares, replies, etc.) on source posts is visible to all users and is widely employed in characterizing the popularity of a given post (Alkhodair et al., 2020; Gao et al., 2021; Yan et al., 2016; Zaman, Fox, & Bradlow, 2014). For example, Alkhodair et al. (2020) and Gao et al. (2021) define the popularity score through the total count of engagements likes, shares, and comments received by the post on Twitter. These engagement attributes are made publicly available via the Twitter API. Gao et al. (2021) showed that the number of reactions a post receives usually grows in early stages (within 24 h of posting) and remains almost constant after a specific period of time (within 10 days of posting), i.e., stable stage. Similar to the previous work, we use the number of shares (E_1), replies (E_2) and reports (E_3) at the stable stage (i.e., at the time we collected the data) as indicators of rumor popularity. More formally, popularity Y_i of a given false rumor X_i is defined as:

$$Y_i = ln[(E_1 + E_2 + E_3^2) + \lambda] \quad (1)$$

where E_1, E_2, and E_3 denote the number of the shares, replies, and reports of the rumor X_i; λ is set to 1 so that the log function always

[6] https://service.account.weibo.com/?type=5&status=4
[7] https://open.weibo.com/

Table 3
Descriptive statistics of the popularity score (Y) in Train, Dev, and Test splits. Y denotes the popularity score of the false rumors.

Popularity score distribution (Y)									
	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, 6)	[6, 7)	[7, 8)	[8, +∞)
All (19,256)									
# of Rumors	5,067	6,144	2,753	1,774	1,274	934	586	372	351
Proportion (%)	26.3	31.9	14.3	9.2	6.6	4.9	3	1.9	1.8
Train (15,404)									
# of Rumors	4,036	4,911	2,225	1,436	1,011	744	462	288	291
Proportion (%)	26.2	31.9	14.4	9.3	6.6	4.8	3.0	1.9	1.9
Dev (1,926)									
# of Rumors	495	651	262	157	132	106	54	39	30
Proportion (%)	25.7	33.9	13.6	8.2	6.9	5.5	2.8	2.0	1.6
Test (1,926)									
# of Rumors	536	582	266	181	132	84	70	45	30
Proportion (%)	27.8	30.2	13.8	9.4	6.9	4.4	3.6	2.3	1.6

Table 4
Descriptive statistics (i.e., Min, Mean, Median, and Max) of the number of tokens in the rumor content (R).

	Min	Mean	Median	Max
All	5	105.2	122	259
Train	5	105.2	122	259
Dev	5	105.3	123	183
Test	7	105.5	124	184

yields a positive value. Note that we give E_3 a higher weight[8] than E_1 and E_2. For a given rumor, we assume that if the fact-checking platform receives more reports, this indicates that the rumor has already received a lot of attention and more users might be unsure about its credibility so they request for it to be fact-checked.

In our initial data exploration, we observed that the number of likes for rumors prior to 2014 was zero, as the 'Like a Post' feature on Weibo was introduced in 2014. For consistency, we do not consider the number of likes when measuring the popularity of rumors given that our dataset contains rumors dating back to 2012. Moreover, the number of views on source posts is another metric that defines popularity scores in social media, especially on YouTube (Kong et al., 2018; Pinto et al., 2013). However, we do not consider it in our paper, as there is no access to the number of views of posts from other users through the Weibo API.

4.4. Data pre-processing

All textual information (i.e., rumor content (R) and user profile description (P)) are pre-processed by removing URLs and user `@mention`. All non-simplified Chinese characters are kept (e.g., traditional Chinese, English, Japanese, etc.) since they appear in the vocabulary list of pre-trained language models (Cui et al., 2020; Sun et al., 2020). The Chinese text is segmented by using the BERT Tokenizer[9] from the HuggingFace library (Wolf et al., 2020). For user attributes (U) (see Table 2), we normalize all numerical variables (e.g., number of friends, followers, statuses, etc.) and transform the Boolean values (e.g., Verified Status (U_7)) into integer values. Note that one can utilize visual information (e.g., images and videos) in the same task. However, we do not consider these features as some rumor cases do not contain these characteristics, or are no longer retrievable.

[8] We believe that some users are unsure about the veracity of suspicious rumors. Therefore, they report them and ask the official Weibo fact-checking platform for fact-checking information (see Fig. 1 for the pipeline of fact-checking on the Weibo platform).

[9] https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertTokenizer

4.5. Dataset description

We remove all false rumors if either the post or the user no longer exist since we need both sources of information for modeling purposes. The final dataset contains 19,256 unique rumors. Each rumor case is linked with the meta-features including (i) rumor engagement attributes (E), (ii) rumor content (R), (iii) user profile description (P) and (iv) user attributes (U). Table 2 displays the categories of meta-features collected via the Weibo API. All rumors and corresponding meta-features will be made publicly available for further investigation by the community.

4.6. Data splits

We random split the rumor dataset into three subsets: (1) train (80%), (2) dev (10%) and (3) test (10%). Tables 3 and 4 display the distribution (i.e., quantity and proportion) of the popularity score and the descriptive statistics (i.e., Mean, Median, and Max) of the number of tokens from the train, development and test sets respectively. The comparison of the statistical distributions of the three subsets showed no significant imbalance. We also notice that around 25% of the false rumors are with low popularity scores, i.e., no or few shares, which demonstrates the importance of identifying false rumors with high likelihood to become popular.

5. Experimental setup

5.1. Predictive models

Since this is the first work on false popularity prediction on Weibo, there is no directly comparable method. Therefore, we opt to evaluate a battery of baseline models to encode textual and user metadata that have been used in previous work on computational misinformation analysis (Alkhodair et al., 2020; Bose, Aletras, Illina, & Fohr, 2022; Rao et al., 2021; Rashkin et al., 2017). Note that one can utilize network information (e.g., rumor propagation network) to model the task. However, we only consider 'out-of-the-box features', such as rumor contents and user-level attributes that are **immediately available** when suspicious rumors are posted on the Weibo platform. This approach allows for the early detection of false rumors with potential high impact in the future.

5.2. Baseline models

To represent textual information, we evaluate four standard baseline models: (1) One-Hot Encoding, (2) Pre-trained Word Vectors (Word2Vec), (3) Graph-based (TextGCN) (Yao, Mao, & Luo, 2019), (4) Transformer-Based Pre-trained Language Models (PLMs). Besides, we also perform experiments on existing rumor detection models. As our goal is the early detection of highly popular false rumors, we have adapted these baseline models by utilizing only readily available features.

SVR+BOW We first employ Support Vector Machine for Regression (SVR) (Cortes & Vapnik, 1995) with an RBF kernel using Bag-of-Words (BOW) weighted using TF-IDF. We use a vocabulary of size 10k most frequent n-grams.

EMB+BiLSTM+ATT (Word2Vec) (Liu & Guo, 2019) We map the text into pre-trained Chinese word embeddings[10] (Li et al., 2018), and then pass them through a bidirectional Long Short-Term Memory network (BiLSTM) (Hochreiter & Schmidhuber, 1997) with a self-attention mechanism. The final weighted representation is then passed through a linear layer for rumor popularity prediction.

TextGCN (Yao et al., 2019) In accordance with Yao et al. (2019), we employ TextGCN to create a graph representation of false rumors by learning the relationships between tokens within the false rumors. We then pass the graph obtained through two layers of graph convolutional networks (GCNs) to generate final predictions.

Pre-trained language models Following Devlin, Chang, Lee, and Toutanova (2019), we directly fine-tune pre-trained transformer-based models on the popularity prediction task by feed [CLS] token representation of the last transformer layer to a linear prediction layer for regression. We evaluate the following models:

- Chinese BERT,[11] pre-trained on the Chinese Wikipedia using character-level tokenization;
- Chinese-BERT-WWM (Cui et al., 2020), an extension of the Chinese BERT model pretrained on larger corpora (e.g., news articles, Baidu Baike, etc.) using the Whole Word Masking (WWM) objective;
- Enhanced Representation Through Knowledge Integration (ERNIE) (Sun et al., 2020), pretrained using both entity-level and phrase-level masking;
- MacBERT (Cui, Che, Liu, Qin, & Yang, 2021), pre-trained using a text correction task with both WWM and n-gram masking methods.

SVR-HF (Source Post + Handcrafted Features) Following Ma, Gao, Wei, Lu, and Wong (2015), we evaluate a SVR model using (i) the rumor content (i.e., source posts) represented with TF-IDF and (ii) a set of handcrafted features obtained from contextual information, such as the ratio of followers to friends and registration time.

Dual-EMO Zhang et al. (2021) uncover the significant role of emotion features from the publisher in detecting online misinformation. To incorporate these findings, we utilize the original pipeline (Zhang et al., 2021) to extract emotional signals, which are combined with information from the source post as inputs for our model.

KG-Trans We also evaluate a knowledge-enhanced rumor detection approach developed by Sun et al. (2023), i.e., the KG-Trans model. This model consists of two key components: (i) extracting entities from the rumor content and (ii) linking the obtained entities in the text modality to entities in an external knowledge graph. Additionally, the source post is encoded using a standard transformer-based encoder, such as BERT, to obtain the post-level representation.

5.3. Developing BERT-Weibo-rumor

Our initial experimental results reveal that vanilla PLMs, such as Chinese BERT, perform better than other encoding methods (e.g., One-Hot, Word2Vec, and Graph), as they have been trained on additional factual information. However, to the best of our knowledge, there are no domain-adaptive PLMs that fit the Weibo platform. For example, BERTweet, a Twitter-adapted PLM, performs better than other vanilla BERT-style models on Twitter datasets.

Following the task adaptive pre-training (Gururangan et al., 2020), we continually pre-train[12] the MacBERT (the one that achieves the best predictive performance in Table 5) on the (1) raw 10 GB Weibo corpus collected using Weibo REST API and; (2) the training set of our specific rumor popularity prediction task. We first train the MacBERT checkpoint on Weibo raw data for one epoch and then further train the MacBERT model on the task-specific training set for 40 epochs. For each epoch, we randomly mask 15% words. We then fine-tune our BERT-Weibo-Rumor model[13] using the same strategy as the original BERT model (Devlin et al., 2019).

5.4. Combining rumor text, user profile description, user attributes and knowledge graph ($R + P + U + KG$)

We also propose a new model (KG-Fusion) that combines rumor content (R), user profile description (P), user attributes (U) and knowledge graph (KG). We first obtain two contextualized representations (i.e., the [CLS] token) H_1 and H_2 for the post itself and the user profile description respectively by passing the text through two transformer-based encoders. Here, we use our 'BERT-Weibo-Rumor' model that achieves the best performance using only the text from the post (see Table 5). The user attributes (U) which are represented by a feature vector are projected into a 128-dimensional representation (H_3).

Knowledge graph PLMs such as BERT capture textual linguistic representations from large-scale corpora (e.g., Wikipedia and books) but lack domain-specific knowledge. In contrast, domain experts (e.g., professional journalists) can use relevant knowledge to reason when reading domain-specific text (e.g., debunking rumors with domain-specific knowledge). Therefore, we also explore the use of an external knowledge graph to enrich the rumor content (R), as introduced in Liu, Zhou, et al. (2020). We first employ a knowledge layer for post-level knowledge queries, which involves matching them with their corresponding triples (Beijing –> Capital –> China) from the knowledge graph (KG), and for knowledge injection, such as extending tokens like 'Beijing' to 'Beijing capital China'. The enriched rumor post is then fed into a transformer encoder[14] to obtain the knowledge representation (H_4). We utilize CN-DBpedia (Xu et al., 2017) as our external knowledge graph. To obtain the final combined representation H_5 of the input, we first project H_1, H_2, H_3, and H_4 into dense vectors of the same dimension and experiment with four different fusion methods:

- We directly concatenate (Concat) the representation of posts (H_1), users' description (H_2), user' profile information (H_3) and knowledge representation H_4 into a single vector (H_5);
- We separately employ a mean pooling layer (Mean Pooling) and a max pooling layer (Max Pooling);
- Finally, we use a self-attention mechanism (Attention) to learn a weighted combination of H_1, H_2, H_3 and H_4.

$$H_5 = H_1 \oplus H_2 \oplus H_3 \oplus H_4 \tag{2}$$

The combined representation (H_5) is finally passed through a fully-connected layer to obtain rumor popularity predictions using a standard mean square error (MSE) loss function.

$$\hat{y} = W * H_5 + b \tag{3}$$

[10] We use 300-dimensional Chinese Word Vectors trained on a Weibo corpus. https://github.com/Embedding/Chinese-Word-Vectors.

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

[12] We use the open source code from Huggingface. https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling.

[13] This model has been released via the HuggingFace platform, which can be reused by the community. https://huggingface.co/YidaM4396/BERT_Weibo_Rumor.

[14] We use a Chinese BERT variant (Liu, Zhou, et al., 2020) that has been pre-trained on WikiZh (i.e., the Chinese Wikipedia corpus) and WebtextZh (i.e., the Chinese question and answer (Q&A) corpus).

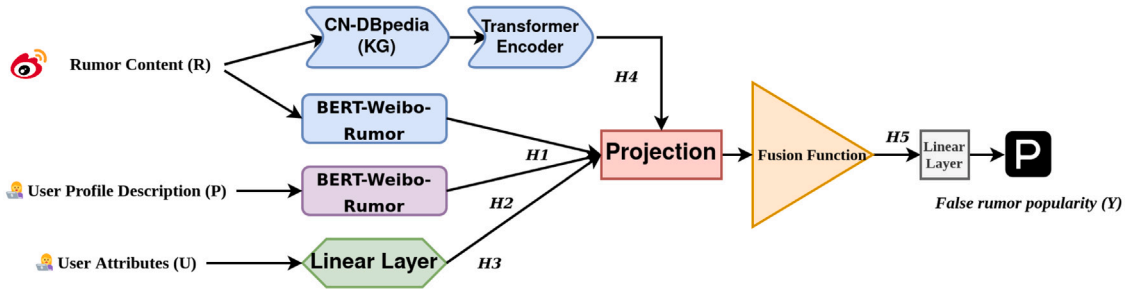

Fig. 2. Combining rumor content and user profile description, user attributes and knowledge graph for rumor popularity prediction. 'Projection' denotes that we project H_1, H_2, H_3 and H_4 into the same dimension. We show the architecture of BERT-Weibo-Rumor in Fig. 3.

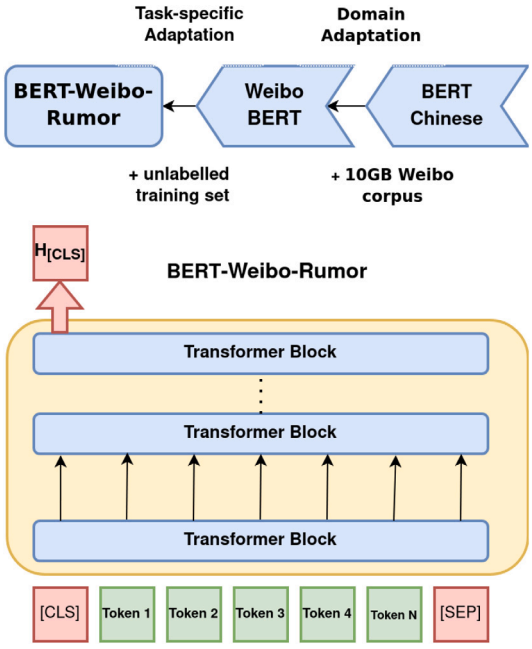

Fig. 3. Model Architecture of BERT-Weibo-Rumor.

Table 5
Average performance (RMSE, Pearson's r, and MAE) for the task of rumor popularity prediction. All Pearson's r values are statistically significant ($p < .001$).

Model	RMSE	Pearson's r	MAE
Weak Baselines			
Mean	1.98 ± 0.0	0.0 ± 0.0	1.56 ± 0.0
Median	2.12 ± 0.0	0.0 ± 0.0	1.44 ± 0.0
Baselines			
SVR	1.62 ± 0.0	0.594 ± 0.0	1.13 ± 0.0
SVR (R+P+U)	1.60 ± 0.01	0.597 ± 0.005	1.12 ± 0.01
EMB+BiLSTM	1.67 ± 0.01	0.539 ± 0.01	1.22 ± 0.02
EMB+BiLSTM (R+P+U)	1.63 ± 0.02	0.565 ± 0.002	1.15 ± 0.03
TextGCN	1.61 ± 0.01	0.595 ± 0.01	1.13 ± 0.02
Chinese BERT	1.59 ± 0.03	0.599 ± 0.019	1.15 ± 0.03
Chinese-BERT-WWM	1.60 ± 0.02	0.598 ± 0.013	1.13 ± 0.01
ERNIE	1.61 ± 0.01	0.585 ± 0.001	1.14 ± 0.02
MacBERT	1.59 ± 0.00	0.603 ± 0.02	1.13 ± 0.02
SVR-HF	1.62 ± 0.03	0.601 ± 0.003	1.13 ± 0.02
Dual-EMO	1.61 ± 0.01	0.594 ± 0.003	1.12 ± 0.01
KG-Trans	1.59 ± 0.02	0.613 ± 0.004	1.14 ± 0.02
BERT-Weibo-Rumor (Ours)	1.57 ± 0.01	0.610 ± 0.02	1.11 ± 0.01
KG-Fusion (Ours)			
Concat ‡	**1.54 ± 0.01**	**0.636 ± 0.004**	**1.08 ± 0.01**
Max Pooling	1.56 ± 0.02	0.626 ± 0.009	1.12 ± 0.01
Mean Pooling	1.57 ± 0.01	0.623 ± 0.002	1.12 ± 0.01
Attention	1.55 ± 0.02	0.623 ± 0.001	1.11 ± 0.01

Note ‡ denotes that the Concat model performs significantly better than all BERT-style models (t-test; $p < .05$).

$$\mathcal{L}_{MSE} = \sum_{i=1}^{D}(y_i - \hat{y}_i)^2 \quad (4)$$

Fig. 2 shows the structure of the proposed neural architecture. For reference, we also conduct experiments on baselines (i.e., SVR and EMB+BiLSTM) around (R+P+U) mixture features.

5.5. Hyperparameters & implementation details

All model hyperparamters are tuned on the development set. We tune the regularization parameter $C \in \{1, 1e1, 1e2, 1e3\}$ and the *ngram* $\in \{(1,1), (1,2), (1,3)\}$ of the SVR, setting $C = 1$ and (1,3). We tune the EMB-BiLSTM *Hidden Size* $\in \{64, 128, 256\}$ and *Dropout* $\in \{0.2, 0.5\}$ with 256 and 0.2 perform best respectively. For all transformer models, we use the 'base' versions with the same architecture and the number of parameters (i.e., 12-layer, 768-dimensional, and 110M model parameters). We fine-tune all transformer based models using the implementations from the HuggingFace library (Wolf et al., 2020). We tune their learning rate range i.e., $lr \in \{2e\text{-}5, 3e\text{-}5, \text{and } 5e\text{-}5\}$ as in Devlin et al. (2019), setting $lr = 5e-5$ for ERNIE, $lr = 3e-5$ for MacBERT and $lr = 2e-5$ for the rest of the models. The input sequence length of the post and user description are set to 256 and 64 covering the maximum length of 99% of all false rumors cases in our dataset (see Table 4). For the fusion network (see Fig. 2), we finetune all the model parameters including the two different BERT encoders for the rumor content (R) and user profile description (P). We use a batch size of 32 for transformer-based models and 128 for the EMB-BiLSTM. All neural networks models are trained by minimizing the Mean Squared Error (MSE) loss using the Adam optimizer (Kingma & Ba, 2015) on a single Nvidia A100 GPU with 40 GB memory.

5.6. Weak baselines

For reference, we also use the mean and median of the popularity scores in the training set as the predicted values of all instances in the test set (i.e., weak baselines).

5.7. Model training and evaluation metrics

We train all of our models three times by performing hyperparameter tuning on the development set using different random seeds. We evaluate model performance using three standard metrics to measure the difference between the actual (y) and predicted ($\hat{y}$) popularity values on the test set. We report the average (mean ± standard deviation across the three runs.

- (i) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

- (ii) Mean Absolutely Error (MAE):

$$MAE = (\frac{1}{n})\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

- (iii) Pearson correlation coefficient (Pearson's r):

$$Pearson's \quad r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \quad (7)$$

5.8. Results

Table 5 shows the results obtained by all models on the false rumor popularity prediction task. We first observe that all models perform substantially better than the two weak baselines (i.e. assigning to all test instances the Mean and Median popularity computed in the training data).

Our proposed neural KG-fusion model (i.e., $R + P + U + KG$ Concat) achieves the lowest RMSE score (1.54), the highest Pearson's r correlation (0.636), and the lowest averaged MAE (1.08) surpassing all the other models. Moreover, the Concat model performs significantly better (t-test; $p < .05$) than the best transformer model 'BERT-Weibo-Rumor' (i.e., RMSE 1.57, Pearson's r 0.610 and MAE=1.11) fine-tuned using only text from the post. This demonstrates that user-related information is complementary to the content of a false rumor for inferring its popularity score. The Concat model performs the best RMSE (1.54), Pearson's r (0.636) and MAE (1.08) than the other three fusion methods. This indicates that the high-dimensional representation obtained by concatenating H_1, H_2, H_3 and H_4 (see Fig. 2) is more informative that the low-dimensional representations from Attention, Max and Mean Pooling.

In general, the majority of the post-level transformer models (i.e., Chinese BERT, Chinese BERT WWM, MacBERT) using the rumor's text as input achieve similar performance with the exception of ERNIE (i.e., RMSE 1.61, Pearson's r 0.585 and MAE 1.14). Our BERT-Weibo-Rumor achieves the RMSE (1.57), Pearson's r (0.610), and MAE (1.11) overall slightly surpassing all other post-level transformer models. Finally, we observe that the two models that are trained from scratch i.e., SVR+BOW (RMSE 1.62, Pearson's r 0.594 and MAE 1.16) and EMB+BiLSTM (RMSE 1.67, Pearson's r 0.539 and MAE 1.14) achieve poorer RMSE and Pearson's r than all transformer-based models except the MAE. These two simpler models have a significantly lower number of parameters and simpler structures than the BERT-style model, suggesting that competitive results can be achieved with models that do not require high computational resources.

6. Analysis

6.1. Ablation study

We perform an ablation study to explore the predictive power of different feature combinations, i.e., rumor content (R), user profile description (P), and user attributes (U). We evaluate five variants: (1) rumor content and user profile description ($R + P$), (2) rumor content and user attributes ($R + U$), (3) user profile description and user attributes ($P + U$), (4) user profile description only (P), (5) user attributes only (U), and (6) $R + P + U$. Besides, we also employ a linear model (i.e., Ridge Regression) to test the predictive performance of each user attribute from U_1 to U_{12} except the 'Verified Status' (i.e., Boolean Value). To make a fair comparison, we test these combinations by using the same experimental setup (i.e., running it three times with different seeds). Table 6 shows the average performance (RMSE, MAE, and Pearson's r).

Table 6
Ablation study. Average performance (RMSE, Pearson's r, and MAE) for the ablation study of rumor popularity prediction. All Pearson's r values are statistically significant ($p < .001$). ‡ We list the results of the KG-Fusion model (i.e., R+P+U+KG) at the top for reference.

Model	RMSE	Pearson's r	MAE
Input			
KG-Fusion ‡	**1.54 ± 0.01**	**0.636 ± 0.004**	**1.08 ± 0.01**
$R + P + U$ **Concat**	1.55 ± 0.01	0.633 ± 0.004	1.10 ± 0.01
$R + P + U$ **Max Pooling**	1.56 ± 0.01	0.625 ± 0.007	1.12 ± 0.02
$R + P + U$ **Mean Pooling**	1.57 ± 0.02	0.621 ± 0.006	1.12 ± 0.02
$R + P + U$ **Attention**	1.56 ± 0.03	0.630 ± 0.004	1.11 ± 0.01
$R + P$	1.56 ± 0.05	0.620 ± 0.005	1.13 ± 0.02
$R + U$	1.56 ± 0.02	0.625 ± 0.004	1.12 ± 0.03
$P + U$	1.76 ± 0.01	0.456 ± 0.005	1.29 ± 0.02
P	1.81 ± 0.01	0.421 ± 0.005	1.33 ± 0.02
U	1.81 ± 0.0	0.470 ± 0.0	1.22 ± 0.0
Linear Regression			
U_1	1.91 ± 0.0	0.278 ± 0.0	1.49 ± 0.0
U_2	1.96 ± 0.0	0.143 ± 0.0	1.52 ± 0.0
U_3	1.92 ± 0.0	0.223 ± 0.0	1.49 ± 0.0
U_4	1.92 ± 0.0	0.309 ± 0.0	1.50 ± 0.0
U_5	1.97 ± 0.0	0.102 ± 0.0	1.55 ± 0.0
U_6	1.98 ± 0.0	0.016 ± 0.0	1.56 ± 0.0
U_8	1.95 ± 0.0	0.168 ± 0.0	1.53 ± 0.0
U_9	1.95 ± 0.0	0.164 ± 0.0	1.53 ± 0.0
U_{10}	1.94 ± 0.0	0.182 ± 0.0	1.52 ± 0.0
U_{11}	1.97 ± 0.0	0.098 ± 0.0	1.55 ± 0.0
U_{12}	1.95 ± 0.0	0.187 ± 0.0	1.52 ± 0.0

We first observe that R+P+U Concat (RMSE 1.55, Pearson's r 0.633 and MAE 1.10) and R+P+U Attention (RMSE 1.56, Pearson's r 0.630 and MAE 1.11) have better predictive performance compared to the BERT-style models that use only R. This suggests that solely rumor content R contains limited information in terms of inferring its future popularity. The remaining three variants (i.e., P+U, P, and U) without using the rumor content (R) perform worse than SVR-BOW and EMB-BiLSTM (see Table 5, suggesting that the textual information of the rumor plays the most important role in predicting its future popularity. Given that the results for all single user attributes used are only just higher than the two weak baseline models (i.e., mean and median), we can infer that individual user attributes are not sufficiently informative in predicting the popularity of false rumors on Wiebo.

Overall, all variants are inferior to the best R+P+U+KG Concat model, which suggests that rumor content and user information are complementary to each other. Finally, linear regression models using individual user attributes U_1-U_2 as input yield results that are close to the mean and median baselines.

6.2. Qualitative analysis of model predictions

To uncover the main limitations of our best model (i.e., $R + P + U$ Concat), we perform an error analysis of false rumor cases where the model predicted a low popularity score for highly popular rumors (Cases Low 1, 2, 3) and vice versa (Cases High 1, 2, 3). Moreover, we analyze two cases where the model correctly predicted a popularity score almost identical to the actual score (Cases Acc 1, 2, 3). These cases are related to the most common topics discussed on Weibo (i.e., 'Politics', 'Social Life' and 'Scientific') (Liu, Zhang, Tu, & Sun, 2015). The false rumor cases (i.e., rumor content R, English translation, and fact-checking hyper-link) together with the actual and predicted popularity scores are shown in Table 7.

Cases low-1, high-1 We first observe that our model has difficulty in accurately predicting the popularity of false rumors related to politics (see Case Low-1 and Case High-3). Given that both cases were posted

Table 7
Examples of prediction actual and predicted popularity scores made by our best performing *R+P+U* Concat model. For each example, we list the original Chinese false rumor *R*, its English Translation, and a link to the corresponding fact-checking page. Note that Weibo requires users to log in to access its fact-checking platform.

	False Rumors Content (in Chinese) and Explanations (in English)	Pred.	Truth	Diff.
False rumors are incorrectly predicted to be low popularity				
	杨澜昨天终于承认了自己的美国籍身份。她理直气壮地说："虽然我入了美国籍，但我出身于中国，所以从原产地角度而言，我不出席美国的两会而出席中国的两会是天经地义的" ...			
Low-1	A false rumor about an official Chinese media host (named '杨澜') has taken U.S. citizenship... **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaJ6wph6Kog	2.76	7.64	-4.88
	六小龄童，昨天因病去世，送"猴哥"最后一程...			
Low-2	A false rumor about the death of a famous actor (named 六小龄童) **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaO6Axk7acl	1.59	8.05	-6.46
	必转！今天一个河北香河的朋友，香河华联超市前拍的照片，这个孩子一看就是被拐卖的！			
Low-3	Please share! This is a photo taken today by a friend in front of a supermarket in Xianghe, Hebei, this child is obviously being trafficked! (This is a false rumor about missing people.) **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaL7Axl66s	4.33	10.12	-5.79
False rumors are incorrectly predicted to be high popularity				
	日本首相安倍晋三正式宣布辞职...			
High-1	Japanese Prime Minister Shinzo Abe officially announced his resignation... **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaN6Qtg66gk	4.36	0.69	3.67
	求助大家帮忙，黄傲雪、7岁、身高，1.2米左右！3月11日中午12:30广州汇豪天下附近 丢失！求大家转发帮忙寻找...			
High-2	Please help, [Girl's Name], 7 years old, about 1.2 meters tall! She disappeared on March 11 at 12:30 p.m. in the [Location]! Please forward to help find this girl... **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaJ6gxc8awl	4.05	0.69	3.36
	#成都提个醒#[太活跃的鱼千万别买] 去买鱼，结果看到摊贩往水盆内加入一种白色粉末，迅速用手搅拌，一会功夫白色粉末溶解，将半死不活的鱼虾倒入其中，一会儿就活蹦乱跳开，仿佛刚从河中捕回来的。这是一种能够致癌的催化剂，俗称鱼浮灵，也对智力有影响。相互转告一下，有必要让更多的人知道！			
High-3	Do not buy live fish that are too active in the supermarket because they contain artificial additives, such as some carcinogens. [Emoji] Share it with each other, it is necessary to let more people know! **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaJ6wtc764d	5.86	2.07	3.79
False rumors that can be accurately predicted by the model				
	一妇女喝了罐饮料，被送进医院，离开了世界。验尸死于於细螺旋体病，追踪她喝的饮料，是直接用罐对嘴饮用。实验证明罐头受到鼠尿感染细螺旋体病毒。			
Acc-1	A woman died after drinking a can of drink contaminated with bacteria carried by rats. **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaJ6wti8qkj	4.56	4.70	-0.14
	Mr.Bean自杀了[泪] I love you, Mr.Bean![伤心][鲜花][蜡烛]			
Acc-2	Mr. Bean committed suicide. [Emoji] I love you Mr.Bean! [Emoji] **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaK6wNk8qwj	4.78	4.64	0.14
	太恐怖了，我以后一定要拔掉！看看！！这个小女孩不幸死于电话充电器，就是因为大人平时充完电后没有及时把充电器插头部位拔出，小女孩拿起充电器另一头来玩含在嘴里不幸 触电身亡..			
Acc-3	Look! This little girl died tragically due to the smartphone charger. Her parents did not unplug the phone from charging, and the little girl picked up the charging head and put it in her mouth to play with it leading to her death. **Fact-checking link:** https://service.account.weibo.com/show?rid=K1CaM6g1f8akh	1.02	1.09	-0.07

by unverified users who have a limited number of followers. However, Case Low-1 eventually became a highly popular false rumor, but Case High-3 received no engagement. This may be due to the fact that the content of the rumor in Case Low-1 is related to populism, which has been on the rise in the past decade on Weibo (Zhang, 2020; Zhang, Liu, & Wen, 2018). Such rumors may attract a lot of engagement by other users.

Cases low-3, high-2 Cases Low-3 and High-2 are both false rumors related to missing people that were published in 2013 and 2014, respectively. In 2012, Professor Yu Jianrong (a famous Chinese sociologist), launched an event called 'Saving children that beg on the streets', calling for people to post and share information about begging children on Weibo to help them find their families (Zhang & Negro, 2013). Since then, Weibo has been widely used to find missing people in case of social emergencies, and many related rumors have emerged (Liu, Uchida, & Utsu, 2020; Shan, Zhao, Wei, & Liu, 2019).

We observe that our model can identify Case Low-3 as a highly popular rumor, but it cannot accurately predict the popularity score (i.e., 10.12). We further explore the corresponding fact-checking link provided in Case Low-3 and discover that the false rumor (with 19k shares and 5k replies) was posted by a famous actor with millions of followers. The first three tokens of the rumor content is '必转!' ('i.e., Please Share!') which reflects the fact that celebrities have high influence on the popularity of false rumors. Moreover, compared against other topics of false rumors such as 'Politics' and 'Science', the missing persons false rumors tend to be more difficult to verify as true or false in a short period of time after they are posted, as they usually require investigation by the authorities.

Cases Acc-1, Acc-3 The last two cases are related to 'Science' and are accurately predicted by our model. We select accurate prediction cases by considering an absolute error less than 0.15, which is the 10-th percentile in the test set. These two cases are quite different (see Fact-checking links in Table 7). Case Acc-1 was posted by a verified user and has become highly popular. On the contrary, case Acc-3 was posted by an average Weibo user and has very low popularity (i.e, no shares and replies). This shows that our best model performs well in learning textual information and user attributes of false rumors about general life knowledge and junk science.

Table 8
Top 5 LIWC features (from Rumor Content (R) and User Profile Description (P)) associated with rumor popularity sorted by Pearson's correlation (r) between the normalized LIWC features frequency and the popularity scores of rumors ($p < 0.001$). We have displayed the top five LIWC categories with the strongest positive and negative correlations, positioned on the left and right sides correspondingly. Note that a positive Pearson's r indicates a positive correlation, and vice versa.

LIWC category	
Rumor content (R)	
conj	affiliation
achieve	family
cause	social
AllPunc	male
OtherP	prep
User profile description (P)	
WC	i
WPS	ppron
work	pronoun
affiliation	female
drives	reward

6.3. Characterizing highly popular false rumors through LIWC

Linguistic analysis Rumors with high popularity typically use stylistic features and sentiment to attract users' attention (Alkhodair et al., 2020). We perform a linguistic analysis to uncover differences in linguistic patterns used between high and low popular false rumors. To this end, we use a standard psycho-linguistic analysis method, i.e. Linguistic Inquiry and Word Count[15] (LIWC) (Pennebaker, Francis, & Booth, 2001), to represent the textual information (i.e., rumor content R and user profile description P) into 95 different psycho-linguistic categories. Table 8 shows the univariate Pearson's correlation test results between the popularity scores of rumors and LIWC features following Schwartz et al. (2013).

Rumor content (R) For the rumor content (R), we observe that LIWC categories such as **Conjunctions** (e.g. and, but, etc.), **Cause** (e.g. because, effect, etc.), and **Achievement** (e.g. win, succeed, better, etc.) are most positively associated with false rumors of high popularity.

In addition, rumor content of false rumors with high popularity contain more punctuation marks, i.e. **AllPunc** (all types of punctuation) and **OtherP** (other uncommon punctuation). We sample some cases with high popularity discovering that punctuation can be used as 'Emoticon' (e.g., ': (', '@_@', ':P') to express emotions. Similarly, these 'Emoticons' have been used to detect Twitter rumors with high engagement rate (Alkhodair et al., 2020).

Besides, we observe some LIWC categories related to emotional expression (e.g. **anger** and **negemo**) are positively correlated with high popularity Weibo rumors. They are also high-frequency words found in false Weibo rumors that can be detected early Song et al. (2019). Note that some common Emoticons (e.g., ':)' and ': (') also belong to the emotion categories in the LIWC dictionary. On the other hand, LIWC categories such as social environment related words (e.g., **social**, **family**, and **male referents** are more common in false rumors with lower popularity scores.

User profile descriptions (P) For user profile descriptions (P), the LIWC categories **Words Count** and **Words per Sentence** show that Weibo users with longer descriptions are likely to share false rumors with high popularity. By exploring descriptive statistics of our dataset, we discover verified users usually have longer descriptions introducing themselves (i.e., average 34 tokens) than unverified users (i.e., average 15 tokens). These verified users (i.e., Verified Status U_7) are also found to have a higher probability of spreading high popularity rumors in the future (see Table 9).

[15] We use a Chinese LIWC version developed by Huang et al. (2012) - https://cliwceg.weebly.com/.

Table 9
Pearson's correlation r between the user attributes (U) and the future popularity of Rumors (p<0.001), sorted in **descending order.** All user attributes are positively correlated with the rumor popularity scores, except for the last one, i.e., **Credit Score** (in bold).

	User attributes (U)
U_7	Verified status
U_1	# of followers
U_4	# of statuses
U_3	# of Bi_Followers
U_{10}	# of replies received
U_{12}	# of all reactions received
U_8	# of shares received
U_9	# of likes received
U_2	# of followees
U_{11}	# of likes received in replies
U_5	# of favorites
U_6	**Credit Score**

The analysis also shows that LIWC categories such as **Work** and **Affiliation** are positively correlated with high popularity rumors, which is the opposite of the negatively correlated LIWC categories discovered in the post. On the other hand, we observe that most negatively correlated LIWC categories (e.g., **i**, **Personal pronouns**, **Total pronouns**) are pronoun-related, however, they do not have high Pearson's r values as they are common words in Weibo user profile descriptions.

User attributes (U) Table 9 displays the sorted Pearson's correlation r between user attributes (from U_1 to U_{12}) and popularity scores of false rumors ($p < 0.001$). We first observe that all user profile attributes are positively correlated with the prevalence of rumors except the 'Credit Score' (U_6). This suggests that the Weibo Credit Score (U_6) is actually a good indicator of user credibility. Thus posts from users with low Credit Scores may need to be prioritized for debunking.

We also observe that the Verified Status' (U_7) of user accounts has the highest Pearson's correlation r, suggesting that false rumors posted by verified Weibo accounts are more likely to receive a larger number of reactions in the future. In social networks, verified accounts are generally considered more credible than average users, and these verified users are significantly more visible in online debates in case of public events (e.g., political events) (González-Bailón & De Domenico, 2021; Hentschel, Alonso, Counts, & Kandylas, 2014). Prior research (Liu et al., 2015) demonstrated that user Verified status (U_7) and number of followers (U_1) can be used as a proxy for the trustworthiness of the Weibo users. Similar to our dataset, high-popularity rumors are more likely to be shared by users with a larger number of followers (U_1) and bi_followers' (U_3) as social media users are more inclined to trust posts that were shared by their friends rather than strangers (Margolin, Hannak, & Weber, 2018). Other positively correlated factors are the reactions including the number of shares (U_8), replies (U_{10}), likes (U_9) that users receive, which reflects the number of interactions they have in the social network.

7. Implications and ethics considerations of our study

In this section, we introduce the ethics considerations, theoretical and practical implications of our research.

7.1. Theoretical implications

The theoretical implications of this work are as follows:

- We define a novel task of predicting future popularity of false rumors which has not been addressed in previous work. We introduce a new direction, and our task can be extended in multilingual and multi-platform settings.
- We provide extensive analyses (see Section 6) including qualitative analysis, psycho-lingual analysis (via LIWC), and user attributes analysis which can be used by social scientists and psychologists to complement studies on analyzing the characteristics of false rumors with high impact (Bronstein, Pennycook, Bear, Rand, & Cannon, 2019; Pennycook & Rand, 2019, 2021).

7.2. Practical implications

We believe that our work has several potential practical implications:

- First, our new dataset (including meta-features), pre-trained language model (i.e., Weibo-BERT-Rumor), and rumor popularity prediction system can be easily re-purposed by fact-checking platforms, professional journalists, researchers, and social media companies. Note that these resources will be released via user-friendly platforms such as HuggingFace and Github.
- Besides, our open source fusion network takes into account both post and user level meta-features to achieve the best predictive performance, and can be used as a strong baseline model for further research.
- Finally, our system can be combined with existing rumor detection systems. For example, in some cases, social media platforms[16] can obtain potential impact immediately upon discovery of a ***false rumor***, which can prevent the spread of high-impact malicious posts at an early stage.

Ethics considerations

Our work complies with Weibo's API policy and has received approval from the Ethics Committee of our institution (Reference Number: 025470). Note that we have also submitted our research proposal to Weibo since we had to apply for the permission for accessing the Weibo official API. All false rumors were debunked and made publicly available by the Weibo fact-checking platform.[17]

Dataset availability

Our new dataset and BERT-Weibo-Rumor model will be made publicly available in compliance with the FAIR principles (i.e., Findable, Accessible, Interoperable and Re-usable) (Wilkinson et al., 2016):

- **Findable & Accessible:** We plan to release the new dataset and model through popular open-source platforms (i.e., Zenodo, Github and Huggingface) with track records.
- **Interoperable:** Our dataset will be released in CSV format including all the features introduced in Table 2, which can be easily imported and processed by most standard data processing tools (e.g, Pandas and NLTK).
- **Re-usable:** Our dataset can be utilized to enhance previous datasets for rumor detection, as we provide the unique post IDs (which can be linked with posts from existing datasets) and additional attributes of each false rumor on Weibo (details see Table 1). Moreover, with the availability of the Transformer library (Wolf et al., 2020), researchers can download and fine-tune the BERT-Weibo-Rumor model directly for other NLP downstream tasks.

[16] Given that most fact-checking platforms (e.g., Weibo rumor debunking platform, PolitiFact, etc.) rely on human resources to manually check the veracity of rumors, with such applications, social media platforms can first address false rumors with higher popularity.

[17] https://service.account.weibo.com/?type=5&status=4

8. Conclusion and future work

This paper presented the first study on future popularity prediction of false rumors on Weibo, based on both post and user-level information. This task is important for the timely detection of high-popularity rumors and complements existing methods for early rumor detection. A key contribution is a new Weibo dataset which includes 19,256 cases of false rumors and their associated popularity score, which is based on the engagement received. To predict the popularity of false rumors, we train a neural model that combines information from the rumor content, user profile description and user attributes, which outperforms strong baselines. Our proposed models and follow-up analysis would enable the prioritization of rumors for moderation and debunking, as well as be beneficial in computational linguistics for analyzing the main characteristics of popular false rumors.

However, we acknowledge that our current contributions, such as our new dataset and the Bert-Weibo-Rumor model, are limited to a mono-lingual setup. Furthermore, we believe that conducting further experiments on feature engineering, such as handcrafting features based on the rumor content, can help improve the model predictive performance. In the future, we plan to extend this work towards studying the popularity of false rumors on different social media platforms and in a multi-lingual setting. Additionally, we plan to conduct a temporal analysis aimed at predicting unseen rumors, following recent work on studying temporal concept drift in computational social science. Hu et al. (2023), Jin, Mu, Maynard, and Bontcheva (2023), Mu, Jin, Bontcheva, and Song (2023).

CRediT authorship contribution statement

Yida Mu: Conceptualization, Resources, Methodology, Software, Validation, Writing – original draft. **Pu Niu:** Funding acquisition, Writing – review & editing, Supervision. **Kalina Bontcheva:** Writing – review & editing, Supervision. **Nikolaos Aletras:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Dataset link: https://doi.org/10.5281/zenodo.8374169.

Acknowledgments

Pu Niu is supported by the Henan Provincial Philosophy and Social Science Planning Fund, China (No. 2023ZT022). This research has been partially supported by the European Union — Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 — Integrating Activities for Advanced Communities" and Grant Agreement n. 871042 ("SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (http://www.sobigdata.eu)). We would like to thank Mali Jin and all reviewers for their valuable feedback.

References

Alkhodair, S. A., Fung, B. C., Ding, S. H., Cheung, W. K., & Huang, S.-C. (2020). Detecting high-engaging breaking news rumors in social media. *ACM Transactions on Management Information Systems (TMIS)*, *12*(1), 1–16.

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating fact checking explanations. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7352–7364).

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2022). Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, *10*, 746–763.

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management*, *58*(4), Article 102569.

Bao, P., Shen, H.-W., Huang, J., & Cheng, X.-Q. (2013). Popularity prediction in microblogging network: A case study on sina weibo. In *Proceedings of the 22nd international conference on world wide web* (pp. 177–178).

Bazmi, P., Asadpour, M., & Shakery, A. (2023). Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Information Processing & Management*, *60*(1), Article 103146.

Bose, T., Aletras, N., Illina, I., & Fohr, D. (2022). Domain classification-based source-specific term penalization for domain adaptation in hate-speech detection. In *Proceedings of the 29th international conference on computational linguistics* (pp. 6656–6666).

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117.

Castillo, C. (2016). *Big crisis data: Social media in disasters and time-critical situations*. Cambridge University Press.

Chen, X., Zhou, F., Zhang, F., & Bonsangue, M. (2021). Catch me if you can: A participant-level rumor detection framework via fine-grained user representation learning. *Information Processing & Management*, *58*(5), Article 102678.

Choi, D., Oh, H., Chun, S., Kwon, T., & Han, J. (2022). Preventing rumor spread with deep learning. *Expert Systems with Applications*, *197*, Article 116688.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: Findings* (pp. 657–668).

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for Chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504–3514.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (Long and Short Papers)* (pp. 4171–4186).

Gao, S., Ma, J., & Chen, Z. (2014). Effective and effortless features for popularity prediction in microblogging network. In *Proceedings of the 23rd international conference on world wide web* (pp. 269–270).

Gao, X., Zheng, Z., Chu, Q., Tang, S., Chen, G., & Deng, Q. (2021). Popularity prediction for single tweet based on heterogeneous bass model. *IEEE Transactions on Knowledge & Data Engineering*, *33*(05), 2165–2178.

Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A., & Chang, S.-F. (2015). Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 907–910).

Glenski, M., Weninger, T., & Volkova, S. (2018). Identifying and understanding user reactions to deceptive and trusted social news sources. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 2: Short Papers)* (pp. 176–181).

González-Bailón, S., & De Domenico, M. (2021). Bots are less central than verified accounts during contentious political events. *Proceedings of the National Academy of Sciences*, *118*(11).

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., et al. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8342–8360).

Hentschel, M., Alonso, O., Counts, S., & Kandylas, V. (2014). Finding users we trust: Scaling up verified Twitter users using their communication patterns. In *Eighth international AAAI conference on weblogs and social media: vol. 8*, (no. 1), (pp. 591–594).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hu, B., Sheng, Q., Cao, J., Zhu, Y., Wang, D., Wang, Z., et al. (2023). Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *Proceedings of the 61st annual meeting of the association for computational linguistics (Volume 5: Industry Track)* (pp. 116–125). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-industry.13, Retrieved from https://aclanthology.org/2023.acl-industry.13.

Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., et al. (2021). Compare to the knowledge: Graph neural fake news detection with external knowledge. 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)* (pp. 754–763).

Huang, C.-L., Chung, C., Hui, N., Lin, Y.-C., Seih, Y.-T., Lam, B., et al. (2012). Development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*, *54*, 185–201.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, *47*(4), 1–38.

Jiang, G., Liu, S., Zhao, Y., Sun, Y., & Zhang, M. (2022). Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, *59*(5), Article 103029.

Jiang, Y., Wang, R., Sun, J., Wang, Y., You, H., & Zhang, Y. (2022). Rumor localization, detection and prediction in social network. *IEEE Transactions on Computational Social Systems*.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 795–816).

Jin, M., Mu, Y., Maynard, D., & Bontcheva, K. (2023). Examining temporal bias in abusive language detection. arXiv preprint arXiv:2309.14146.

Karmakharm, T., Aletras, N., & Bontcheva, K. (2019). Journalist-in-the-loop: Continuous learning as a service for rumour analysis. 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): System demonstrations* (pp. 115–120).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International conference on learning representations*. Retrieved from http://arxiv.org/abs/1412.6980.

Kochkina, E., Hossain, T., Logan IV, R. L., Arana-Catania, M., Procter, R., Zubiaga, A., et al. (2023). Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, *60*(1), Article 103116.

Kong, Q., Rizoiu, M.-A., Wu, S., & Xie, L. (2018). Will this video go viral: Explaining and predicting the popularity of youtube videos. In *Companion proceedings of the the web conference 2018* (pp. 175–178).

Lampos, V., Aletras, N., Preoţiuc-Pietro, D., & Cohn, T. (2014). Predicting and characterising user impact on Twitter. In *14th Conference of the European chapter of the association for computational linguistics* (pp. 405–413).

Li, L., Situ, R., Gao, J., Yang, Z., & Liu, W. (2017). A hybrid model combining convolutional neural network with xgboost for predicting social media popularity. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 1912–1917).

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 2: Short papers)* (pp. 138–143).

Lin, H., Ma, J., Cheng, M., Yang, Z., Chen, L., & Chen, G. (2021). Rumor detection on Twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10035–10047).

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, *337*, 325–338.

Liu, Y., Uchida, O., & Utsu, K. (2020). A proposal on disaster information and rescue request sharing application using sina weibo. In *2020 5th International conference on computer and communication systems* (pp. 419–423). IEEE.

Liu, Z., Zhang, L., Tu, C., & Sun, M. (2015). Statistical and semantic analysis of rumors in chinese social media. *Scientia Sinica Informationis*, *45*(12), 1536–1546.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., et al. (2020). K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence: vol. 34*, (no. 03), (pp. 2901–2908).

Lu, H.-y., Fan, C., Song, X., & Fang, W. (2021). A novel few-shot learning based multi-modality fusion model for COVID-19 rumor detection from online social media. *PeerJ Computer Science*, *7*, Article e688.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., et al. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th international joint conference on artificial intelligence* (pp. 3818–3824).

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1751–1754).

Ma, J., Gao, W., & Wong, K.-F. (2017). 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)* (pp. 708–717).

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, *35*(2), 196–219.

McParlane, P. J., Moshfeghi, Y., & Jose, J. M. (2014). "Nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of international conference on multimedia retrieval* (pp. 385–391).

Middleton, S. E., Middleton, L., & Modafferi, S. (2013). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17.

Mu, Y., & Aletras, N. (2020). Identifying Twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, *6*, Article e325.

Mu, Y., Bontcheva, K., & Aletras, N. (2023). It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the association for computational linguistics: EACL 2023* (pp. 724–731).

Mu, Y., Jin, M., Bontcheva, K., & Song, X. (2023). Examining temporalities on stance detection towards COVID-19 vaccination. arXiv preprint arXiv:2304.04806.

Mu, Y., Niu, P., & Aletras, N. (2022). Identifying and characterizing active citizens who refute misinformation in social media. In *14th ACM Web science conference* (pp. 401–410).

Mu, Y., Song, X., Bontcheva, K., & Aletras, N. (2023). Examining the limitations of computational rumor detection models trained on static datasets. arXiv preprint arXiv:2309.11576.

Nobre, G. P., Ferreira, C. H., & Almeida, J. M. (2022). A hierarchical network-oriented analysis of user participation in misinformation spread on WhatsApp. *Information Processing & Management*, *59*(1), Article 102757.

Parikh, S. B., Patil, V., Makawana, R., & Atrey, P. K. (2019). Towards impact scoring of fake news. In *2019 IEEE conference on multimedia information processing and retrieval* (pp. 529–533). IEEE.

Pavleska, T., Školkay, A., Zankova, B., Ribeiro, N., & Bechmann, A. (2018). Performance analysis of fact-checking organizations and initiatives in Europe: A critical overview of online platforms fighting fake news. *Social Media and Convergence*, *29*.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402.

Pinto, H., Almeida, J. M., & Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the 6th ACM international conference on web search and data mining* (pp. 365–374).

Piotrkowicz, A., Dimitrova, V., Otterbacher, J., & Markert, K. (2017). Headlines matter: Using headlines to predict the popularity of news articles on Twitter and Facebook. In *Proceedings of the international AAAI conference on web and social media, Vol. 11, no. 1* (pp. 656–659).

Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining* (pp. 518–527). IEEE.

Rao, D., Miao, X., Jiang, Z., & Li, R. (2021). STANKER: Stacking network based on level-grained attention-masked bert for rumor detection on social media. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3347–3363).

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, *8*(9), Article e73791.

Shan, S., Zhao, F., Wei, Y., & Liu, M. (2019). Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—A case study of Weibo (Chinese Twitter). *Safety Science*, *115*, 393–413.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 395–405).

Silva, A., Han, Y., Luo, L., Karunasekera, S., & Leckie, C. (2021). Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, *58*(5), Article 102618.

Smith, J. H., & Bastian, N. D. (2022). A ranked solution for social media fact checking using epidemic spread modeling. *Information Sciences*, *589*, 550–563.

Song, C., Yang, C., Chen, H., Tu, C., Liu, Z., & Sun, M. (2019). CED: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3035–3047.

Song, C., Yang, C., Chen, H., Tu, C., Liu, Z., & Sun, M. (2021). CED: Credible early detection of social media rumors. *IEEE Transactions on Knowledge & Data Engineering*, *33*(08), 3035–3047.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., et al. (2020). Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence: vol. 34*, (no. 05), (pp. 8968–8975).

Sun, M., Zhang, X., Ma, J., Xie, S., Liu, Y., & Philip, S. Y. (2023). Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). 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)* (pp. 809–819).

Trzciński, T., & Rokita, P. (2017). Predicting popularity of online videos using support vector regression. *IEEE Transactions on Multimedia*, *19*(11), 2561–2570.

Vo, N., & Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 275–284).

Vo, N., & Lee, K. (2019). Learning from fact-checkers: Analysis and generation of fact-checking language. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 335–344).

Vo, N., & Lee, K. (2020). Where are the facts? Searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 7717–7731).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wan, P., Wang, X., Pang, G., Wang, L., & Min, G. (2023). A novel rumor detection with multi-objective loss functions in online social networks. *Expert Systems with Applications*, *213*, Article 119239.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 2: Short Papers)* (pp. 422–426).

Wang, Wang, T., Ye, X., Zhu, J., & Lee, J. (2016). Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm. *Sustainability*, *8*(1), 25.

Wei, L., Hu, D., Zhou, W., Yue, Z., & Hu, S. (2021). Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection. 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)* (pp. 3845–3854).

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on web search and data mining* (pp. 261–270).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6, Retrieved from https://aclanthology.org/2020.emnlp-demos.6.

Xia, R., Xuan, K., & Yu, J. (2020). A state-independent and time-evolving network with applications to early rumor detection. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 9042–9051).

Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., et al. (2017). CN-DBpedia: A never-ending Chinese knowledge extraction system. In *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 428–438). Springer.

Yan, Y., Tan, Z., Gao, X., Tang, S., & Chen, G. (2016). STH-Bass: A spatial-temporal heterogeneous bass model to predict single-tweet popularity. In *International conference on database systems for advanced applications* (pp. 18–32). Springer.

Yang, Y., Wang, Y., Wang, L., & Meng, J. (2022). PostCom2DR: Utilizing information from post and comments to detect rumors. *Expert Systems with Applications*, *189*, Article 116071.

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence: vol. 33*, (no. 01), (pp. 7370–7377).

Yuan, C., Ma, Q., Zhou, W., Han, J., & Hu, S. (2020). Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In *Proceedings of the 28th international conference on computational linguistics* (pp. 5444–5454).

Zaman, T., Fox, E. B., & Bradlow, E. T. (2014). A bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, *8*(3), 1583–1611.

Zhang, D. (2020). Digital nationalism on weibo on the 70th Chinese national day. *The Journal of Communication and Media Studies*, *6*(1), 1–19.

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021* (pp. 3465–3476).

Zhang, Y., Liu, J., & Wen, J.-R. (2018). Nationalism on Weibo: Towards a multifaceted understanding of Chinese nationalism. *The China Quarterly*, *235*, 758–783.

Zhang, Z., & Negro, G. (2013). Weibo in China: Understanding its development through communication analysis and cultural studies. *Communication, Politics & Culture*, *46*(2), 199–216.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., & Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1513–1522).

Zhou, H., Ma, T., Rong, H., Qian, Y., Tian, Y., & Al-Nabhan, N. (2022). MDMN: Multi-task and domain adaptation based multi-modal network for early rumor detection. *Expert Systems with Applications*, *195*, Article 116517.

Zhou, K., Shu, C., Li, B., & Lau, J. H. (2019). Early rumour detection. 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)* (pp. 1614–1623).

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, *51*(2), 1–36.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One*, *11*(3), Article e0150989.