Drivers of True and False Information Spread: A Causal Study of User Sharing Behaviors

Ling Sun[1,2(✉)], Kathleen M. Carley[2], and Yuan Rao[1]

[1] Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China
sunling@stu.xjtu.edu.cn, raoyuan@mail.xjtu.edu.cn
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
{lings,carley}@andrew.cmu.edu

Abstract. Analyzing and predicting user information-sharing behavior on online social platforms is a crucial task in social sciences. While current prediction tasks primarily emphasize accuracy, they often neglect the underlying motivations that drive user behavior, hindering a fundamental understanding and control of the information spreading environment. To address this, we analyze and quantify potential factors that may drive user sharing behavior based on social theories. Our limited derived feature set achieves over 85% accuracy in predicting user behavior on two real-world datasets, demonstrating its effectiveness. Notably, through employing causal inference techniques, our analysis on true and false information spread reveals that users with lower authority are more susceptible to being misled by false information. In contrast, the propagation of truthful news is often driven by personal preference or influenced by users' social circles. By uncovering these underlying motivations, our approach facilitates a deeper comprehension of the online information ecosystem, contributing to more effective management strategies for false information mitigation.

Keywords: Behavior Prediction · False Information · Causal Inference

1 Introduction

Online social media enables individuals to obtain information in a cheap and handy way, yet it also promotes the spread of misinformation. Predicting and analyzing users' information sharing behavior, in environments where true and false information coexist, is a crucial task in the field of information governance.

Current user behavior prediction work primarily focuses on utilizing machine learning or deep learning methods to improve prediction accuracy, but largely overlooks the underlying driving factors behind the behavior, resulting in low interpretability and credibility of the results, making it difficult to apply in real-world scenarios. Furthermore, current research on misinformation dissemination mainly focuses on comparing the spread patterns of true and false information

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 174–183, 2024.
https://doi.org/10.1007/978-3-031-72241-7_17

from a macro perspective. They have found that false information tends to spread faster, deeper, and more broadly than real information [16,18], and pointed out there are differences in novelty, topic distribution, and sentiment distribution between true and false information. However, their analyses are relatively independent, overlooking the interplay between features, and the research primarily concentrates on positive samples, i.e., users who participated in sharing, without combining and contrasting with negative samples, i.e., users who were exposed to the news but did not share it, making it difficult to uncover the true driving factors behind users' behavior. Due to the numerous and interrelated factors determining user behavior, accurately predicting user behavior and analyzing the underlying reasons is a challenging task.

Given the limitations of current research, we integrate the discovery of motivations behind information spreading with the task of predicting user behavior. From both theoretical and practical perspectives, we provide a more detailed explanation on how various factors drive user behavior. Specifically, our approach draws upon social theories to identify and extract the most crucial driving factors from complex social data, and achieves high accuracy in behavior prediction through a very limited feature set. Moreover, instead of relying on feature importance rankings from the prediction task, we introduce causal graphs to describe the interactions between features, through cross-analysis of bot accounts, human accounts, and the propagation of true and false information, we unveil the key underlying reasons genuinely influencing user behavior. In summary, this paper:

- Identifies and quantifies the potential factors influencing whether users share a news post based on reliable social theories, and validates the effectiveness of the derived features through user behavior prediction task.
- Constructs a valid causal graph to intuitively illustrate the relationships between various factors, and uncovers the motivations of users in the propagation of fake and real news through cross causal analysis and comparison.

2 Related Works

2.1 User Sharing Behavior Prediction

The user sharing behavior prediction task aims to predict whether a user will share a specific news posts based on relevant features. Researchers have proposed various machine learning-based prediction methods that integrate multimodal data and different types of features. Zhang et al. [17] developed an attention-based convolutional neural network based on the content being shared. Firdaus et al. [7]comprehensively modeled users' past tweets and sharing behavior, analyzing interest, sentiment, and personality traits of users to predict the likelihood of sharing. Sun et al. [14] employed sequential hypergraph neural networks and attention mechanisms to model time-varying user preference and predict the next infected user in information propagation.

However, the aforementioned methods primarily aim to improve prediction accuracy, making it challenging to interpret the results from a social science

perspective, thereby limiting their practical application. Recently, Sun et al. [15] designed a causal-enriched deep attention (CEDA) framework to evaluate the causal effects of input variables on retweet behaviors during prediction, improving the interpretability of model.

2.2 Information Propagation Analysis

Many researchers have attempted to measure and analyze the prevalence of false information on social media. Vosoughi et al. [16] found that falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all fields. Zhou et al. [18] concluded four patterns for false information propagation, namely, More-Spreader, Further-Distance, Stronger-Engagement, and Denser-Network.

Several studies further delved into investigating the underlying reasons behind the viral propagation of false information. Based on social theories, the factors that motivate users to spread information can be reflected in four aspects: 1. News attributes, such as news topic, sentiment, source, etc. News topics influence the writing style of posts and the sharing tendency of users, and often invoke emotional responses [10]. Therefore, news with specific attributes may attract more attention from users; 2. User attributes, such as gender, age, number of friends, activity level, authority level, etc. For instance, Altay et al. [1] found that users with more friends want to keep a positive self-image of themselves, so they share less fake news; 3. User interest, users typically follow what they like [4], and prefer information that confirms their preexisting attitudes [6]. Therefore, user interest is a key determinant of information sharing, yet it is influenced by numerous factors, the interests expressed in user posts may not reflect genuine preferences, but rather stem from the echo chamber effect, which limits exposure to diverse information [4]; 4. Social influence, social identity theory reveals that users tend to conform to the viewpoints prevalent within their community to gain acceptance and achieve a sense of belonging [3]. Gimpel et al. [9] found that fake news shared by users is often shared by their trusted family and friends.

Cheng et al. [4] classified user sharing behavior into intentional and unintentional to identify suspicious users. Bui et al. [3]further analyzed the influence of factors such as social identity and news polarity on sharing intentions. In this study, we provide a more detailed explanation on how various factors drive propagation by integrating user behavior prediction with causal inference technology.

3 Analysis and Calculation of Sharing Driving Factors

Our experiments and analysis based on the publicly available fake news detection dataset, FakeNewsNet [13], which comprises news data linked to two fact-checking platforms: GossipCop[1] and PolitiFact[2]. The PolitiFact dataset predominantly addresses political topics, whereas GossipCop focuses primarily on entertainment news. These datasets encompass comprehensive information, including

[1] https://www.gossipcop.com/.
[2] https://www.politifact.com/.

news text, sources, user profiles and their historical behaviors. Based on social theories, we categorize potential factors influencing user sharing behavior into four categories, and utilize additional computations and tools to enhance more valuable features, details are provided in Table 1.

News Attributes. Textual features of news posts, such as topics and sentiments, have been shown to be closely related to their diffusion effects [3,16]. Meanwhile, the source website and the number of engagements indicate the credibility and popularity of news, which are also valuable. Thus, we identify news sentiments from the content using the Google Cloud Natural Language API[3], then obtain the source ratings from the Web of Trust API[4]. The total number of tweets, retweets, and comments for each news are used to measure its popularity.

Table 1. Features included, calculation methods/tools used, and range of values.

	Metrics	Description	Formulation/Tool	Scale
News Attributes	Source	The source website for publishing the news	–	–
	Source score	Security score of the website	WOT API	[0,1]
	Popularity	# tweets, retweets and replies	#tweets+#retweets +#replies	–
	Sentiment	Sentiment of news content	Google Cloud Natural Language API	[−1,1]
	Topic	Topic of news content	Google Cloud Natural Language API	–
User Attributes	Basic attributes	'user_id', 'created_at', '#favorites', '#friends', '#listed', '#followers', '#statuses', 'verified'	–	–
	Activity	User's activity level	Norm(0.5 *(#statuses/(time - created_at))+ 0.5*((#favourites + #friends)/2))	[0,1]
	Authority	User's authority level	Norm((#followers +#listed)/(1+#friends)* (1.2 if verified else 0.8))	[0,1]
	Bot score	Probability of being a bot	BotHunter	[0,1]
	Negativity	Proportion of posts with negative sentiments	Vader API	[0,1]
	Emotional	Proportion of posts with strong sentiments	Vader API	[0,1]
User Interest	Interest score	User's interest in news	SimCSE	[0,1]
Social Influence	Neighbor influence	Number and influence of users' neighbors who shared the news before	–	

User Attributes. Users' behavior is significantly influenced by their selection bias, which are closely related to their attributes. Research [4] found that users' verification status, status and friend count are associated with the probability of being suspicious. Therefore, we utilize the user profiles as basic attributes, and calculate the activity and authority scores based on user behavioral data. Since

[3] https://cloud.google.com/natural-language.
[4] https://www.mywot.com/.

numerous bot accounts exist on social platforms, we employ the BotHunter [2] to estimate the probability of an account being a bot. Furthermore, as emotional arousal is a crucial factor driving information sharing [10], we leverage the Vader API [11] to identify the sentiment of each post published by users, and calculate the proportion of negative sentiment (¡ 0) posts as the user's negativity score and the proportion of extreme sentiment (¿ 0.5 or ¡ –0.5) posts as their emotional score.

User Interest. User interests are influenced by various factors, such as the interests of their friends and the biases of recommendation algorithms [4]. Therefore, we treat interest scores as independent from user attributes and calculate them separately. Specifically, since typical language models like BERT [5] are unsuitable for computing similarity between short texts, we employ the SimCSE, an improved method based on contrastive learning [8]. To reduce computational costs, we concatenate all tweets of a user into a single document and design a sliding window, computing the similarity between each window and the target news text, and obtain the final interest score by averaging these similarity scores.

Social Influence. User behavior is easily influenced by their friends or family [3]. However, networks on social platforms are very large and sparse. To retain the most valuable influences, we construct a directed user network based on their historical behaviors, where an edge exists only between users who have directly or indirectly shared each other's posts, and calculate the social influence exerted on a user by quantifying the influence of their neighbors.

4 User Sharing Behavior Prediction

4.1 Data Sampling and Experimental Settings

Based on the constructed user interaction network, we sample negative instances from the neighbors of known positive instances at a 1:1 ratio. To ensure that the negative instances had the potential to encounter the news, we only sample from nodes that have previously received information from the positive instances. To compare the behavioral differences between human and bot accounts, we classified accounts with a bot score greater than 0.6 as bot accounts. Conversely, accounts with a bot score less than 0.4 were defined as human users. The final statistics is shown in the Table 2. We split the data into training and test sets with a 7:3 ratio and conducted baseline experiments using multiple machine learning classifiers, including Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), and Decision Trees (DT). The results demonstrated that RF achieved the best predictive performance. Consequently, all experiments and analyses presented in this work are based on the RF classifier.

Table 2. Data statistics

Dataset	News	Positive samples	Negtive samples	Bots	Human
Politifact-fake	341	276,748	231,011	162,846	121,432
Politifact-real	240	311,923	250,733	163,276	172,500
Gossipcop-fake	3430	965,641	826,866	466,798	613,531
Gossipcop-real	6903	812,788	698,528	641,471	452,280

4.2 Feature Importance Analysis

As shown in Table 3, the experimental results indicate that fully utilizing all features always leads to the highest accuracy, highlighting the importance of integrating diverse features. Besides, user attributes performs better than other features in predicting users' tendencies to share both real and fake news, particularly in Gossipcop dataset, where the model attains an accuracy of 85.90%. This approves that user behavior is significantly influenced by their own preferences. Social influence also plays a pivotal role, especially for fake news in Gossipcop dataset. This could be because entertainment news is less contentious than political news, so people rely more on social engagements and are easily deceived by fake news with specific characteristics, rather than out of interest.

Table 3. Prediction results with different features.

Features	Politifact						Gossipcop					
	All		Fake		Real		All		Fake		Real	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
All	**86.44**	**86.74**	**86.65**	**86.83**	**86.26**	**86.64**	**89.97**	**90.31**	**88.89**	**89.41**	**91.42**	**91.56**
News atts.	54.90	70.77	54.30	70.28	55.31	71.18	53.83	69.00	59.74	73.12	51.39	65.58
User atts.	77.82	78.08	77.16	77.18	76.30	77.34	85.90	86.52	83.03	84.02	88.06	88.26
Interest	58.11	59.28	59.04	58.26	57.77	62.00	57.48	59.13	58.43	60.76	57.46	56.84
Social inf.	62.86	67.49	62.49	65.98	63.30	68.81	60.78	72.07	62.33	72.47	58.87	70.86

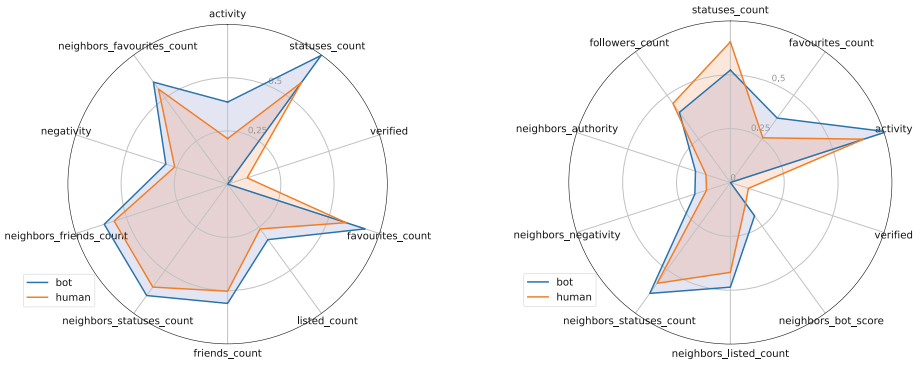
4.3 Bots and Human Behavior Analysis

Given the differing behavior patterns and driving factors between bot accounts and human accounts, we extracted potential bot and human accounts from the dataset and analyze the distinctions between them. We first extract the ten most distinguishing features and visualize them in Fig. 1. Notably, across both datasets, human accounts exhibit a higher proportion of being verified, while bot accounts display more active.

Additionally, we conducted separate predictions, results are presented in Table 4. It can be observed that the prediction accuracy for bot accounts is higher than for human accounts across both datasets. Notably, the accuracy for bot accounts is about 1.6% and 5% higher than human in the Politact and Gossipcap datasets, respectively, and the F1 score for real news in Gossipcop is even 27.7% higher than that for human accounts. This indicates that while bot accounts may mimic human-like features, they exhibit different behavioral pattern, which is simpler and more predictable.

Table 4. Prediction results with bots and human accounts.

Identity	Politifact						Gossipcop					
	All		Fake		Real		All		Fake		Real	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Bots	**82.01**	**74.45**	**81.71**	**74.14**	**82.25**	**75.60**	**89.56**	**89.96**	**84.59**	**80.16**	**93.18**	**94.53**
Human	80.46	70.44	80.37	65.28	80.62	73.25	84.57	70.06	83.34	71.68	86.36	67.33



(a) Politifact (b) Gossipcop

Fig. 1. Visualization of the key features of bots and human accounts

5 Motivations Discovery

Given that classifiers can only identify the correlation between features and behavior, rather than causal relationships, and often neglect the interactions between features, we constructed a causal graph based on social science theories, and utilized causal intervention strategy to calculate the causal effect of each feature on user behavior, and ultimately proved that the driving factors exhibit significant differences across various dissemination scenarios.

5.1 Causal Graph Construction

Based on social science theories, we first construct a causal graph for four factors and the results: user behavior, as illustrated in the Fig. 2. As we illustrated in Sect. 3, all four factors can directly influence user behavior. Additionally, based on the echo chamber effect, user attributes and social influence may affect user interest. For instance, users are more likely to encounter information shared by those around them and content recommended by algorithms based on their attributes [1,4]. Furthermore, the social influence a user experiences is simultaneously affected by both user attributes and news attributes, that is, users with more friends generally experience greater social influence, and users may share highly popular news based on social conformity theory [3].

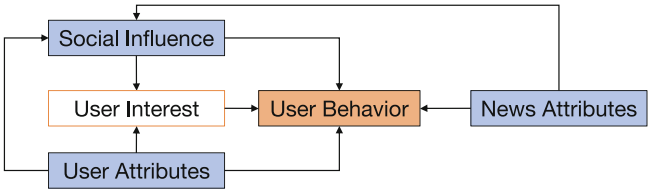
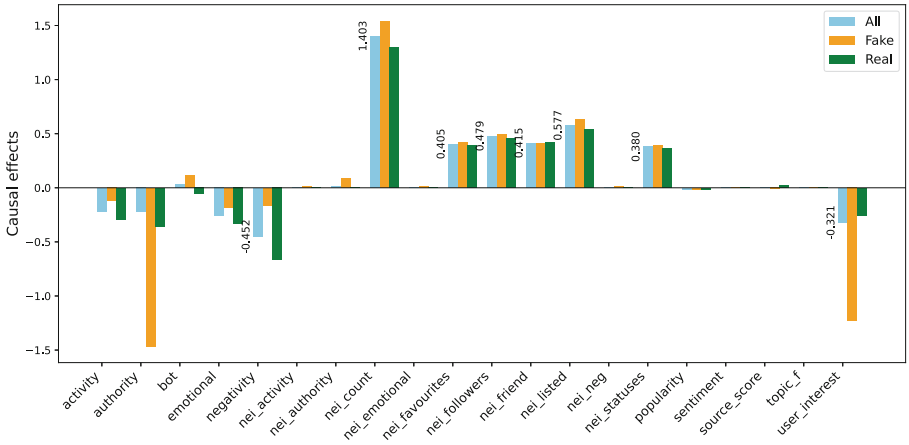
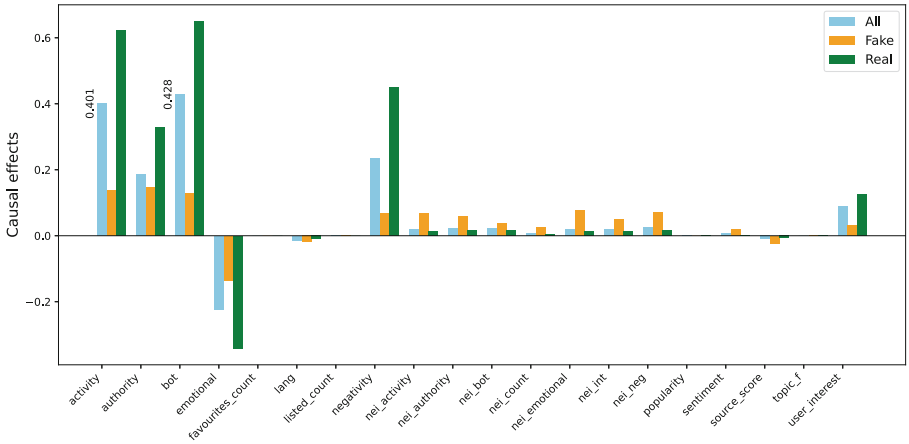


Fig. 2. The causal relationship between four types of features and user behavior



(a) Politifact



(b) Gossipcop

Fig. 3. The causal effects of features on user behavior (top 20 significant)

5.2 Causal Calculation and Motivation Discovery

Based on the causal graph, we aim to estimate the causal effect of each treatment (feature) on outcome(user behavior) using the DoWhy [12] tool. Given a Treatment (X), and an Outcome (Y), the estimated effect can be calculated by intervening the value of X:

$$\mathbb{E}[Y|do(X = x^{'})] - \mathbb{E}[Y|do(X = x)] \tag{1}$$

Specifically, for a given feature X (e.g., "User Interest"), we first utilize the backdoor criterion to identify potential confounders, which are variables that simultaneously affect both the treatment variable and the outcome variable "User Behavior". According to the backdoor criterion, we need to control for the variables "News Attributes" and "Social Influence" to block backdoor paths and eliminate confounding effects. Subsequently, we estimate the causal effect using Linear Regression (LR) due to its high computational efficiency and ease of interpretation. The coefficient of the LR directly represents the causal effects of the treatment on the outcome.

We computed the causal effects of all features on user behavior for both datasets, as illustrated in Fig. 3. For the PolitiFact dataset, it can be observed that the authority of users has the most significant impact on the dissemination of fake news, showing a pronounced negative correlation –1.48. This indicates that users with lower authority are more likely to be deceived by fake news. In contrast, for the dissemination of real news, the number of neighbors who have shared the news plays the most crucial positive role (1.32), suggesting that user behavior is largely driven by social conformity.

The causal distribution in Gossipcop differs significantly from PolitiFact. User attributes such as activity, authority, bot score, and negativity all exhibit relatively strong positive correlations with sharing behavior, while the influence from neighbors is comparatively lower. This indicates that for entertainment news, user behavior is more influenced by personal characteristics and interests. Moreover, we observe instances where bot accounts actively engage in retweeting truthful news, likely as a strategy to enhance their authority. These findings demonstrate that news propagation drivers vary across different topics, suggesting that diverse measures may be needed in information control.

6 Discussion and Conclusion

To better understand the driving mechanisms behind information propagation and facilitate control of the online information environment, we analyze and quantify the underlying reasons for user information sharing behavior from multiple aspects, grounded in social theories. Through user behavior prediction experiments and causal analysis, we demonstrate the effectiveness of our extracted features. We find that although bot accounts mimic human features, their behavioral patterns are still distinguishable and more predictable. Moreover, the veracity and topic of information lead to different distributions of driving factors, suggesting that distinct strategies should be employed in various scenarios.

References

1. Altay, S., Hacquin, A., Mercier, H.: Why do so few people share fake news? it hurts their reputation. New Media Soc. **24**(6), 1303–1324 (2022)
2. Beskow, D.M., Carley, K.M.: Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In: SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, vol. 3 (2018)
3. Bui, Q.N., Moriuchi, E.: Sharing intention of politicized news on social media: mediators and moderators. In: 57th Hawaii International Conference on System Sciences, HICSS 2024, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, 3–6 January 2024, pp. 6086–6095 (2024)
4. Cheng, L., Guo, R., Shu, K., Liu, H.: Causal understanding of fake news dissemination on social media. In: KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, 14–18 August 2021, pp. 148–157 (2021)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, vol. 1, pp. 4171–4186 (2019)
6. Dmj, L., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
7. Firdaus, S.N., Ding, C., Sadeghian, A.: Retweet prediction based on topic, emotion and personality. Online Soc. Netw. Media **25**, 100165 (2021)
8. Gao, T., Yao, X., Chen, D.: Simcse: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual, 7–11 November 2021, pp. 6894–6910 (2021)
9. Gimpel, H., Heger, S., Olenberger, C., Utz, L.: The effectiveness of social norms in fighting fake news on social media. J. Manag. Inf. Syst. **38**(1), 196–221 (2021)
10. Horner, C.G., Galletta, D.F., Crawford, J., Shirsat, A.: Emotions: the unexplored fuel of fake news on social media. J. Manag. Inf. Syst. **38**(4), 1039–1066 (2021)
11. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: The Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, 1–4 June 2014 (2014)
12. Sharma, A., Kiciman, E.: Dowhy: an end-to-end library for causal inference. CoRR arxiv:2011.04216 (2020)
13. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data **8**(3), 171–188 (2020)
14. Sun, L., Rao, Y., Zhang, X., Lan, Y.: MS-HGAT: memory-enhanced sequential hypergraph attention network for information diffusion prediction. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, pp. 4156–4164 (2022)
15. Sun, W., Liu, X.F.: Deep attention framework for retweet prediction enriched with causal inferences. Appl. Intell. **53**(20), 24293–24313 (2023)
16. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
17. Zhang, Q., Gong, Y., Wu, J., Huang, H., Huang, X.: Retweet prediction with attention-based deep neural network. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, 24–28 October 2016, pp. 75–84 (2016)
18. Zhou, X., Zafarani, R.: Network-based fake news detection: a pattern-driven approach. SIGKDD Explor. **21**(2), 48–60 (2019)