Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Information Processing and Management 59 (2022) 102935

Contents lists available at ScienceDirect

Information Processing and Management

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

Investigation of the determinants for misinformation correction effectiveness on social media during COVID-19 pandemic

Yuqi Zhang, Bin Guo *, Yasan Ding, Jiaqi Liu, Chen Qiu, Sicong Liu, Zhiwen Yu

Northwestern Polytechnical University, Xi'an, 710129 CN, China

ARTICLE INFO

Keywords:
Misinformation
Correction
Cognitive factor
Social media
Data-driven
Microblog

ABSTRACT

The rapid dissemination of misinformation in social media during the COVID-19 pandemic triggers panic and threatens the pandemic preparedness and control. Correction is a crucial countermeasure to debunk misperceptions. However, the effective mechanism of correction on social media is not fully verified. Previous works focus on psychological theories and experimental studies, while the applicability of conclusions to the actual social media is unclear. This study explores determinants governing the effectiveness of misinformation corrections on social media with a combination of a data-driven approach and related theories on psychology and communication. Specifically, referring to the Backfire Effect, Source Credibility, and Audience's role in dissemination theories, we propose five hypotheses containing seven potential factors (regarding correction content and publishers' influence), e.g., the proportion of original misinformation and warnings of misinformation. Then, we obtain 1487 significant COVID-19 related corrections on Microblog between January 1st, 2020 and April 30th, 2020, and conduct annotations, which characterize each piece of correction based on the aforementioned factors. We demonstrate several promising conclusions through a comprehensive analysis of the dataset. For example, mentioning excessive original misinformation in corrections would not undermine people's believability within a short period after reading; warnings of misinformation in a demanding tone make correction worse; determinants of correction effectiveness vary among different topics of misinformation. Finally, we build a regression model to predict correction effectiveness. These results provide practical suggestions on misinformation correction on social media, and a tool to guide practitioners to revise corrections before publishing, leading to ideal efficacies.

1. Introduction

The world has suffered severe attacks from 'infodemic' on social media during the COVID-19 pandemic, including rumors, stigma, and conspiracy theories (we refer to such inaccurate content collectively as "misinformation" below) (Guo, Ding, Yao, Liang, & Yu, 2020; Lederer, 2020; Luo, Xue, & Hu, 2020; Pian, Chi, & Ma, 2021). The misinformation spreads fear and stigma to the public and undermines the adoption of reasonable preventions and policies for control (Bridgman et al., 2020; Zhou, Xiu, Wang, & Yu, 2021), that poses threats to pandemic preparedness and control. Therefore, mitigating the severe impacts of misinformation has aroused widespread attention. Researchers have tried to combat them with the methods like fact-checking, correcting, and debunking misinformation (Burel, Farrell, & Alani, 2021; Vraga & Bode, 2020). Correction on social media is one of the crucial

* Corresponding author.
E-mail addresses: yuqizhang@mail.nwpu.edu.cn (Y. Zhang), guobin.keio@gmail.com (B. Guo), yasanding@mail.nwpu.edu.cn (Y. Ding), jqliu@nwpu.edu.cn (J. Liu), qiuchen@nwpu.edu.cn (C. Qiu), scliu@nwpu.edu.cn (S. Liu), zhiwenyu@nwpu.edu.cn (Z. Yu).

https://doi.org/10.1016/j.ipm.2022.102935
Received 10 December 2021; Received in revised form 10 February 2022; Accepted 21 March 2022
Available online 5 April 2022

countermeasures for online misinformation, that debunks a false claim or a misperception through posts published by users (Vraga & Bode, 2020). However, the mechanism of misinformation correction is not clear on social media, for example, how to organize the content appropriately and how the publishers affect the efficacy of corrections are not well resolved. Therefore, it is necessary to further explore the instinct mechanisms of misinformation correction on social media and guide practitioners to correct it effectively.

Previous works (Budak, Agrawal, & El Abbadi, 2011; He, Song, Chen, & Jiang, 2012; Nguyen, Yan, & Thai, 2013; Saxena et al., 2020; Tong & Wu, 2018) mainly study the effectiveness of misinformation corrections on social media from two perspectives: "truth campaigns" and cognitive effectiveness of corrections. "Truth campaign" explores the promotion of corrections dissemination, mostly based on diffusion cascades and greedy algorithms. Cognitive effectiveness of corrections studies the cognitive effects of corrections on individuals, with the aim to find effective ways to convey corrections (Dai, Yu, & Shen, 2021; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Rich & Zaragoza, 2020). This paper focuses on the latter perspective. Cognitive effectiveness of misinformation correction mainly focuses on psychology and communication theories and experiments. According to Lewandowsky et al. (2012), the characteristics of an effective debunking post are summarized grounded in psychological theories (e.g., a convincing explanation, a concise statement). Moreover, much experimental research is conducted to examine the influencing factors of the effectiveness of corrections, e.g., the timeliness of debunking misinformation (Rich & Zaragoza, 2020), message order (before misinformation or after misinformation) and debiasing message (including in corrections or not) (Dai et al., 2021). Although these works have been done by pioneering research on the cognitive effects of misinformation correction, there still exists the following constraints: first, the data samples are usually small because of the cost of social experiments, so the representativeness of the result can be affected. Second, surveyed or interviewed data are often objective and can easily be influenced by "observer bias". Third, the applicability of these works to social media is unclear, because their results are not obtained from the actual data flow on social media.

In this paper, we explore the determinants governing the effectiveness of misinformation corrections based on the comprehensive analysis of a COVID-19 online dataset and propose a regression model to predict correction effectiveness, given the correction post's basic features. Our dataset consists of 1487 corrections on significant COVID-19 related events between January 1st, 2020 and April 30th, 2020, that is retrieved from Microblog on Internet. Furthermore, our work is grounded in theories on psychology and communication (e.g., Backfire Effect, Source Credibility, Disconfirmation Bias) (Kim & Dennis, 2018; Lewandowsky et al., 2012; Nyhan & Reifler, 2010). To our best knowledge, our work is the first to explore the cognitive effectiveness of misinformation corrections based on the data-driven approach and theories on psychology and communication. Besides, we fill the gap in the lack of qualitative mechanisms to guide effective corrections.

According to theories on psychology and communication (i.e., Backfire Effect, Source Credibility, and Audience's Roles in Dissemination) (Ecker, Lewandowsky, & Apai, 2011; Ecker, Lewandowsky, & Tang, 2010; Johnson & Seifert, 1994; Kim & Dennis, 2018; Shu, Bernard, & Liu, 2019; Van den Broek, Young, Tzeng, Linderholm, et al., 1999), we propose five hypotheses, that assume the associations between seven potential factors and the effectiveness of misinformation correction. Afterward, based on COVID-19 related corrections from Microblog, we evaluate the effectiveness of corrections using their social contexts and examine the relations in assumptions. Through the comprehensive analysis, we obtain five findings, i.e., mentioning excessive original misinformation in corrections would not undermine the believability of people within a short period after reading; warnings of misinformation in a demanding tone make corrections worse; concise corrections are more effective; persuasive graphic explanations are appealed to be valued; influential media should take more responsibility. According to our findings, we appeal for the corrections which are short, concise, persuasive, rich in graphics, in a gentle tone, and published by influential users. In addition, we demonstrate that the effects of the above factors on corrections vary among different topics. Finally, we build the regression model to predict the effectiveness of corrections. The R^2 (Goodness of Fit) of the optimal model in the training dataset is up to 0.66 and up to 0.46 in the testing dataset. The proposed model obtains the key relations among these variables. Our novel contributions are summarized as follow:

- As far as we know, our work is the first time that studies the cognitive effectiveness of misinformation corrections with the combination of a data-driven approach and related theories on psychology and communication. We propose assumptions based on theories (i.e., Backfire Effect, Mental Model of Misinformation, Source Credibility, and Audience's Roles in Dissemination) and conduct extensive evaluations on 1487 corrections on influential COVID-19 related events from January 1st, 2020, to April 30th, 2020, on Microblog.
- Based upon a comprehensive analysis of results, we obtain the determinants of the efficacy of corrections and provide explanations for results according to "Continued Effect of Misinformation" (Johnson & Seifert, 1994; Wilkes & Leatherbarrow, 1988) and "Disconfirmation Bias" theories. We also propose five suggestions on the good practice of effective corrections on social media.
- We build an effective regression model to predict the efficacies of corrections. This model captures the key relations among features and efficiencies of corrections. It can guide practitioners to revise their corrections before publishing and achieve ideal efficiency.

The rest of the paper is organized as follows. We introduce related research about misinformation correction in Section 2. Then, we elaborate on the problem in Section 3, and describe the proposed methodology in Section 4. Subsequently, the evaluation results are elucidated in Section 5 and the implications are discussed in Section 6, and we discuss limitations and future work in Section 7. Finally, we comprehensively provide our conclusions in Section 8.

2. Literature review

Existing works mainly study the effectiveness of misinformation corrections from two perspectives: "truth campaigns" and the cognitive effectiveness of corrections. In this section, we summarize recent works in these two fields.

2.1. Truth campaigns

Misinformation intervention strategies can be divided into two categories: influence blocking and truth campaigns. Influence blocking strategy functions by blocking critical nodes or edges in the spread of misinformation to minimize the flow of misinformation (Wu & Pan, 2017; Yang, Liao, Shen, Cheng, & Chen, 2018). And truth campaigns strategy select the optimal seed nodes to diffuse the truth information to influence users' awareness and reduce the proliferation of misinformation (Zareie & Sakellariou, 2021). The latter strategy is related to the focus of this paper: the dissemination of misinformation correction, so the methods of truth campaigns are elaborated here. In truth campaigns, the misinformation and truth both spread in the network. It chooses the top-k originators to diffuse the corrective message and reverses viewpoints of nodes already affected by false claims. The ultimate goal is to minimize the number of users adopting the misinformation in the network. Nguyen et al. (2013) stimulated the diffusion process of the truth based on the IC or LT diffusion model. They designed a greedy algorithm to select the best set of seed users to start the spread of the corresponding fact, making sure at least the β-fraction of users in the network can be decontaminated. Yang, Li, and Giua (2020) set two different thresholds for each individual in the diffusion model: influence threshold and decision threshold. The influence threshold took effects in activating a node receiving no information, and the decision threshold was used while convincing a node to switch its perceptions. Tong and Wu (2018) assumed that more than one truth campaign may attempt to intervene in the spread of misinformation and they considered multiple diffusion cascades with different cascade priorities in the dissemination model.

The aforementioned methods are simply based on the structural information of the network and ignore the users' perceptions and realistic situations. Recent works address this problem. Song, Hsu, and Lee (2017) found out the influence of misinformation is time-efficient, and after a specific time point, it can barely affect people. On top of this, they proposed a top-k debunkers identification algorithm, finding the nodes which influence the maximum number of users before a given deadline. Saxena et al. (2020) proposed an opinion model that took the fluctuation of users' opinions into account and designed a mitigation solution to select a subset of debunkers to maximize the number of users who turned over the wrong pre-beliefs. Zhang, Yang, and Du (2021) considered another realistic situation, where both seed nodes and boost nodes existed in the diffusion model. The boost nodes could be more likely to adopt the corrective message when they receive it.

The above studies improve the effects of corrections from the perspective of network structure. Although some of them take users' opinions and realistic conditions into consideration, the associations between factors in corrections and correction effectiveness are not estimated. For example, the content of corrections and the influence of debunkers can significantly affect the acceptance of people on misinformation corrections. It is necessary to understand the mechanisms of correction effectiveness and helps in mitigating the negative impacts of misinformation.

2.2. Cognitive effectiveness of corrections

Misinformation is usually persistent and difficult to entirely be clarified and can influence users' perceptions even after corrections, namely, the continued effect of misinformation (Andrea & Radvansky, 2020; Desai & Reimers, 2019; Johnson & Seifert, 1994). Many kinds of research have confirmed it. For example, Ecker et al. (2011) found out the retraction on misinformation rarely has the intended effect of eliminating the reliance on misperceptions even when people acknowledge and remember this retraction. The reason that results in debunking misinformation such difficulty has been explored in many studies in psychology. One possible reason is motivated reasoning. Individuals can protect their pre-existing attitudes and prefer to receive the information they believed before (Jerit & Barabas, 2012; Taber & Lodge, 2006). When people with wrong beliefs encounter a corrective message, they may feel be challenged and refuse to accept the fact. In addition, misinformation tends to be sticky and persistent in the brain due to the mental models people build for the process of misinformation (Thorson, 2016). Another reason is that corrections can sometimes lead to backfire effects (Lewandowsky et al., 2012), meaning that misinformation correction cannot achieve the intended effect but strengthen the misperception (e.g., familiarity backfire effect, the overkill backfire effect).

During the COVID-19 pandemic, a large proportion of misinformation crowds in the social media. The persuasion strategies of misinformation-containing posts (Chen, Xiao, & Mao, 2021) and users' poor ability to discriminate misinformation (Zrnec, Poženel, & Lavbič, 2022) nudge the viral dissemination of misinformation, threatening the pandemic preparedness and control. Correction is an important intervention for misinformation on social media. The effects of corrections are vital to ensure the orderly epidemic control. Therefore, it is crucial to understand the mechanism how the correction can affect the perceptions of individuals effectively, which is named as cognitive effectiveness of correction. Some efforts have been taken to study the effective mechanism of correcting misinformation on psychology and communication. Lewandowsky et al. (2012) summarized the characteristics of an effective debunking post grounded in psychological theories (e.g., a convincing explanation, a concise statement). Bode and Vraga (2018) investigated the self-correction ability of Facebook through a web-based survey of participants. The results demonstrated that corrections recommended by the algorithm and corrections received from social connections are equally effective in reducing misperceptions. Ecker, Hogan, and Lewandowsky (2017) confirmed the familiarity backfire effect of misinformation. They found that the corrections repeating a piece of misinformation can lead to a stronger reduction of misperceptions than those

without misinformation repetition. But another study demonstrated that corrections mentioning misinformation would decrease the audience's belief in falsehoods, who had not been exposed to misinformation before corrections (Ecker, O'Reilly, Reid, & Chang, 2020). Ecker, Lewandowsky, Jayawardana, and Mladenovic (2019) tested the overkill effect and found evidence against it that the correction with a larger number of counterarguments led to as much or more misperceptions reduction compared to one with a smaller number of counterarguments. Rich and Zaragoza (2020) assessed the influence of the timeliness of debunking misinformation on the efficacy of corrections. They also found in experiments that people's belief in misinformation can increase over two days after correction. Dai et al. (2021) investigated two factors that could affect the effectiveness of correction, namely, message order (before misinformation or after misinformation) and debiasing message (including in corrections or not). Through online experiments, they revealed that misinformation corrections are most effective when they are conveyed after misinformation and include debiasing messages. Bautista, Zhang, and Gwizdka (2021) conducted interviews on healthcare professionals to build the conceptual model of these professionals' act of correcting health misinformation on social media, to guide effective corrections for authorities.

These works on the efficacy of misinformation correction explore how the content and format of correction influence its effectiveness. Some are theoretical research, and the others are experiment research based on surveys or interviews. This ''well-designed'' research paradigm is suitable for exploring specific questions deeply. But they do not capture the associations from the real data flow on social media, the universality of conclusions is unclear. Besides, it has constraints in experiment cost and small samples. In this work, we apply a data-driven approach to explore the factors governing correction effectiveness.

3. Problem statement

The negative influence of misinformation has aroused dramatic attention in society during the COVID-19 pandemic and many interventions have been implemented, like fact-checking, correcting. Correction on social media is one of the crucial countermeasures for online misinformation. However, the effective mechanism of misinformation correction is not clear on social media, for example, how effective mainstream corrections are, how to organize the content appropriately, and how the credibility of publishers affect the efficacy of corrections are not solved.

In this work, we propose assumptions about the determinants governing the effectiveness of misinformation corrections according to ''continued effects of misinformation'' and ''backfire effects'' theories. These hypotheses include seven factors that are potentially associated with the effectiveness of corrections, i.e., the proportion of original misinformation, length of the post, textual warning of misinformation, graphic warning of misinformation, explanation, graphic explanation, and influence of publisher. Besides, according to the theory that users play different roles in disseminating misinformation, it raises our curiosity about the role of users in the spread of corrections.

- Familiarity Backfire Effect

Repetition of one message makes people familiar with the information and rarely suspect its veracity. That is because prior knowledge smooths the process of thinking (Schwarz, Sanna, Skurnik, & Yoon, 2007). In other words, repetition of information strengthens the familiarity and builds up people's belief in it. Unfortunately, in corrections of misinformation, it is unavoidable to mention misinformation. Otherwise, corrections cannot convey key points needed to be clarified. Consequently, corrections mentioning too much original misinformation make people familiar with it and likely to believe it, contrary to the original intentions. This effect is named as familiarity backfire effect. Considering that the mainstream correction posts on Microblog cannot avoid mentioning original misinformation, we hypothesize:

H1: The proportion of original misinformation in posts decreases the correction effectiveness.

- Overkill Backfire Effect

According to the research (Schwarz et al., 2007), information is more likely to be accepted if it is easy to process. However, for misinformation corrections, more evidence means more successful. However, it turns out that fewer and simple arguments are more effective in reducing misperceptions. That is because too many arguments can take lots of effort to understand and simple statements can be more attractive cognitively. Therefore, the overkill backfire effect means that providing many arguments sometimes cannot reduce wrong beliefs and even reinforce wrong beliefs. It is necessary that keep corrections clean and easy to process. Considering the complexity of the corrections in Microblog varies (here we take the length of the post as a measurement of complexity), we assume:

H2: If the length of corrections is too long, the effects of corrections will be greatly affected.

- Explanations Filled the Gaps in Mental Model

When people hear misinformation, they would build a mental model with this misinformation. This mental model records the associations among the key elements in a message, e.g., causality, correlativity. Once the false information is refuted, some key elements or associations in the mental model could be overthrown and there could be gaps left. According to existing research (Ecker et al., 2011, 2010; Johnson & Seifert, 1994; Van den Broek et al., 1999), if corrections do not provide alternative explanations to fill those gaps, people would continue to use that inaccurate information. It is the reason that they may be uncomfortable with gaps in their knowledge and prefer a complete model even it is inaccurate. Moreover, this effect was verified by experiments. For

example, in an experiment, people read a piece of misinformation that a warehouse fire was made of paint and gas cans along with explosions (Johnson & Seifert, 1994; Seifert, 2002; Wilkes & Leatherbarrow, 1988). And in people's mental model, it could be like "paints and gas cans led to explosions" and "explosions caused the fire in a warehouse". Then people were told that paint and cans were not present at the fire, and when asked questions about the cause of the fire, they invoked the paint and cans despite having just acknowledged these things were not present at the fire. An alternative explanation that some accelerants contributed to the fire was provided for people. After that, people were less likely to mention "paint and cans". Therefore, it is crucial to provide alternative explanations to replace the misperceptions. Compelling explanations can be why the misinformation is wrong or why the originator of misinformation disseminated the false information. Graphics are significantly more effective than text in reducing misperception (Alda et al., 2012), we hypothesize accordingly:

H3: The alternative explanations improve the effectiveness of corrections, especially for the graphic explanations.

- Explicit Warnings

Although people usually expect that the information they encounter is valid, misinformation is unavoidable. With explicit warnings appearing in front of misinformation, it can induce a temporary state of skepticism, which can improve people's ability to discriminate between truth and original misinformation. Besides, Ecker et al. (2010) investigated whether explicit warnings can reduce the continued influence of misinformation. It turned out that a specific warning giving details about the continued influence effect can reduce the misperceptions and a particular warning combined with alternative explanations can reduce misperception more effectively. The correction posts on Microblog usually use textual alerts like "xx is wrong!" and graphic warnings. We want to explore whether these kinds of warnings positively affect the correction effectiveness.

H4: Textual and graphic warnings help in promoting the correction effectiveness.

- Source Credibility

When people meet a piece of information, they usually leverage their prior knowledge. The prior knowledge impacts how people evaluate the trustworthiness of the content (Hovland, Janis, & Kelley, 1953). When people consider the veracity of a message, one kind of crucial prior knowledge they rely on is the credibility of the source. They prefer to believe information from sources they trust rather than doubtful ones (McCracken, 1989). Kim and Dennis (2018) studied whether the presentation format of fake news affects people's believability. They found out that high ratings of the source have a significant and positive effect on believability. Therefore, credible sources make people more likely to accept and believe corrections, thus improving the correction effectiveness. In our work, we evaluate the credibility of publishers based on their influence.

H5: The influential publisher makes corrections more acceptable.

- Various Roles of Users in the Dissemination

During the spread of fake news, individuals play different roles. Persuaders spread fake news with supporting opinions to convince others. Clarifiers are someone who proposes skeptical and opposing opinions to clarify fake news. Gullible users have a low ability to distinguish between true and false information and are easily persuaded to accept false information (Shu et al., 2019). Analogously, we assume that users in the dissemination of corrections on social media also have various roles, and thereby we question:

RQ1: What are the effects of various users on correction effectiveness in the dissemination of corrections?

4. Method

In the method, as Fig. 1 shows, after assumptions, we obtain correction posts and their social interactions, characterize the correction posts based on factors from hypothesis, and conduct correlation analysis and exploratory analysis on the dataset. On top of analysis results, we examine the validity of assumptions and build the regression model to predict correction effectiveness.

In this section, firstly, we describe our dataset concerning 1487 correction posts from Microblog. Subsequently, we demonstrate details on dataset labeling and evaluation of the effectiveness of corrections. Finally, we illustrate the implementations of regression models.

4.1. Online data collection

We select the Sina Weibo platform (which is also known as Microblog) as our research target. Sina Weibo is one of the largest social media in China. Users can publish the posts and interact with other users, e.g., leaving comments, liking and retweeting to a post. "Weibo Pi Yao" is an official account to report the misinformation and publish corrections, which is operated by the authorities of the Sina Weibo platform. We acquire all posts published by the account from January 1st, 2020, to April 30th, 2020. The total number of collected posts is 1608 and 1487 correction posts are reserved after removing non-correction posts. Then the social interactions of reserved posts are collected. From Sina Weibo, we collect 1487 correction posts from 532 unique publishers from January 1st, 2020, to April 30th, 2020. We also gather 70,772 comments, 52,682 retweet users, and 51,006 comment users of these posts. The scripts of data collection are implemented in the Python Requests package.

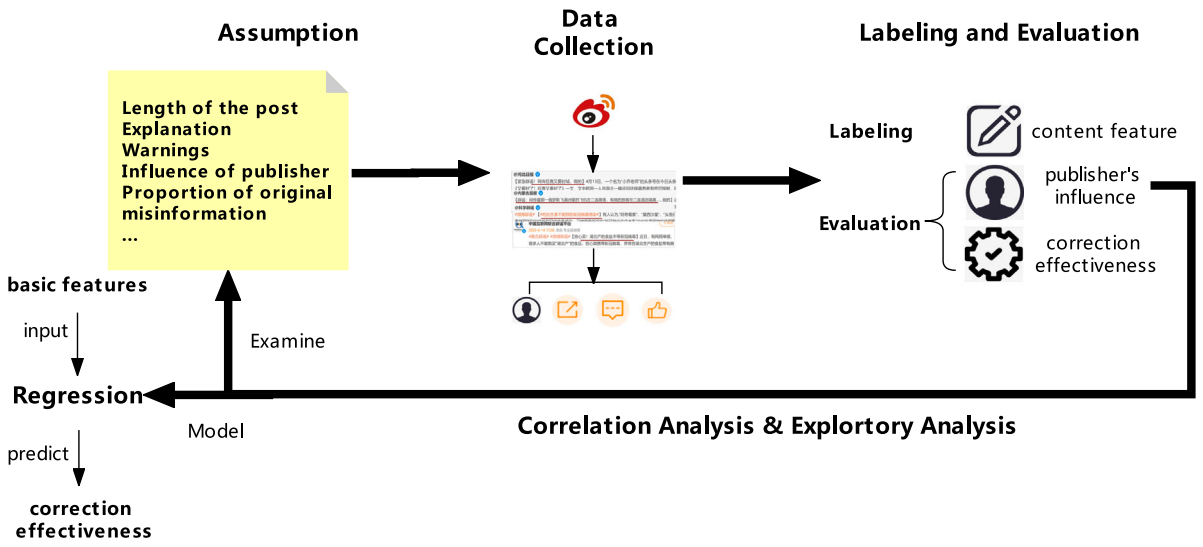


Fig. 1. Illustration of methodology.

Ethics statement. Approval and informed confirmation are not needed. We only obtain publicly available data for the data collection which anyone can access, and any content with privacy restrictions is not collected.

Posts of correction. We collect 1487 correction posts to the influential misinformation on Microblog from January 1st, 2020, to April 30th, 2020. This collection includes the post texts and their metadata, i.e., the unique identity number of the post, the publisher of the post, the publish timestamp, retweet counts, like counts and comment counts. During data preprocessing, we remove the special strings of digital elements (i.e., URLs, emoticons, "@xxx") and the hashtag("#xxx#") is reserved because we assume that it can convey some considerable information.

Through the exploratory analysis of the collected posts, we summarize the topics of posts into four main categories: "control measures in the epidemic", "situation of the epidemic", "medical knowledge", and "supply and safety". The numbers of corrections in each category are: 357,759,171, and 120. The rest eighty posts are classified as "other" type. "Control measures in the epidemic" type usually relates to some control measures, such as transportation control, lockdown, resumption of work, return to school, vaccines, etc. "Situation of epidemic" type is associated with new cases of COVID-19, virus transmission, isolation to potential virus carriers, etc. "Medical knowledge" type contains corrections on the common sense of medicine, virus prevention, virus self-test, etc. In supply and safety, there are correction posts about medical supplies, food safety and so on. It should be mentioned that there are some posts including multiple corrections, and the topic category of this kind posts is determined by the majority. The definition and examples are shown in Table 1.

Comments of posts. Evaluation of the effectiveness of correction in Section 4.2 is based on the sentiment analysis on posts' comments. Therefore, we gather the comments of the pre-collected posts, and due to the limits of API, we only obtain comments of every post in the first 50 pages. 70,772 comments are collected finally. Each piece of data consists of the comment text, the publisher of the comment, the timestamp, like counts.

User information. According to the hypothesis in Section 3, the influence of users is also considered as a possible influencing factor, so we collect the information about publishers (i.e., the number of followers and influence data from module "influence of yesterday" in Microblog, details shown in Section 4.2) to evaluate their influence. Here we obtain data from 535 unique publishers. Besides, for further exploration about the effect of users in the dissemination of corrections, we gather the fan number of users interacting with pre-collected posts. Specific analysis is shown in Section 5. Finally, we collect information about 52,682 retweet users and 51,006 comment users.

4.2. Label influencing factors and evaluation

According to our hypotheses, we summarize seven potential influencing factors: the proportion of original misinformation, length of the post, textual warning of misinformation, graphic warning of misinformation, explanation, graphic explanation, and influence of publisher. Then we characterize the correction posts based on these factors. Six factors related to content can be labeled manually, and the influence of the publisher is evaluated with the user information gathered in Section 4.1. Finally, we evaluate the effectiveness of corrections based on the social behaviors of the posts and sentiment analysis of comments.

Table 1
Description of collected data.

Category	Definition	Example of the correction post
Control measures in the epidemic	Relates to some control measures, such as transportation control, lockdown, resumption of work, return to school, vaccines, etc.	"On 9th March, it is widely spread online that XX University released a notice to inform senior students to return to school for graduation in June. Today, the college claims this message is fabricated. They mention that the school returning will be arranged in the near future."
Situation of the epidemic	Be associated with new cases of COVID-19, virus transmission, isolation to potential virus carriers, etc.	"On 11th March, someone spread the message 'one person was diagnosed with COVID-19 infection', we solemnly declare that the news is a rumor. And we have reported to the police, the originator will take the legal responsibility for this."
Medical knowledge	Common senses of medicine, virus prevention, virus self-test, etc.	"Recently there is a new online that taking a sip of water every 15 min to keep the throat moist can prevent the virus. According to specialists, there is no relationship between virus infection and dry throat, and drinking too much water may bring extra strain to the body."
Supply and safety	About medical supplies, food safety and so on.	"China is a major producer of masks in the world, and its annual export volume remains stable at more than 70% of its production scale. China has never issued a ban on the export of masks and their raw materials, and enterprises can carry out trades by market-oriented principles. said Li Xinggan, director-general of the Department of Foreign Trade of the Ministry of Commerce."

Content feature manual labeling. We recruit three undergraduate students to label the dataset, well-educated college students aged from 20 to 22 years old. Before the annotation, we explain the annotation standard to them last for half an hour. And during the labeling, we answer their questions about the rules. The standard of annotation is: The proportion of original misinformation is calculated by the length of original misinformation mentioned in the post divided by the length of the post. Textual warning of misinformation is labeled as to whether the post provides text warning before first mentioning the fake message. The graphic warning is labeled as to whether the post has warnings in the form of pictures. And the explanation, referring to theories in Section 3, is characterized as whether the post explains why the misinformation is wrong or why originators of misinformation disseminated the false information. The graphic explanation is annotated as to whether the correction includes the pictures of the reason. Some posts contain more than one piece of correction. For this kind of correction, the labeling rules for "length of the post", "the proportion of original misinformation", and "influence of publisher" stay the same. The labeling result of other factors is calculated as the number of corrections satisfying the requirements of the feature divided by the total number of corrections mentioned in the post. The category of correction is classified based on the topics described in Section 4.1. The details of annotations are presented in Table 2. During data annotations, disagreement among annotators may occur, and thus we elaborately make rules to handle conflicts for each feature. For most features that are easier to be identified and to reach an agreement (e.g., category, textual warning, graphic warning, graphic explanation, and explanation), we use the majority of annotations as the final label. For features that are harder to reach an agreement (e.g., the proportion of original misinformation, and length of the post), we take the average of all annotations as the final label.

Publisher's influence evaluation. Although the number of followers is a common metric to evaluate the influence of users, it is not a reliable metric due to the markets of zombie fans.[1] Therefore, we collect user influence data from a module called "influence of yesterday", maintained by Microblog officials. Although the number of followers is a common metric to evaluate the influence of users, it is not a reliable metric due to the markets of zombie fans. Therefore, we collect user influence data from a module called "influence of yesterday", maintained by Microblog officials. The module contains three elements: the number of posts published yesterday, the number of being read yesterday, and the number of interactions with users yesterday. These characteristics combined with followers can approximately model the real influence of users. The influence of publisher i, IN_i is calculated by the weighted sum of the average number of being read yesterday $\frac{R_i}{P_i}$, the average number of interactions yesterday $\frac{I_i}{P_i}$, and the number of followers F_i. R_i is the number of being read yesterday, P_i is the number of the posts published yesterday, and I_i represents the number of interactions yesterday. We assumed that numerous readings and interactions of a post greatly reflect the publisher's influence. Therefore, we take $\alpha = 0.8, \beta = 0.2$ here.

$$IN_i = \alpha * \left(\frac{R_i + I_i}{P_i} \right) + \beta * F_i \tag{1}$$

Evaluation on effectiveness of correction. Evaluating the effectiveness of correction posts is a prerequisite, to model the relationship between potential factors and correction effectiveness. According to work by Kim and Dennis (2018), they found out that the believability of information positively influences social behaviors such as liking, retweeting, and commenting positively. In other

[1] Zombie fans usually are bot accounts generated by the platform and some can be normal users. They can interact with employers' generated content (e.g., like, retweet). Some users can buy some zombie fans to build up their influence (Fuyong, Jing, Qianqian, & Xufeng, 2012).

Table 2
Labeling and evaluation of collected data.

Factor	Definition	Annotation standard
Proportion of original misinformation	The percentage of textual misinformation mentioned in correction text	Be calculated by the length of original misinformation mentioned in the post divided by the length of the post.
Length of the post	The length of the text	The number of characters of the post which excludes the special strings (e.g., URLs, emoticons, "@xxx") and punctuations.
Explanation	Whether the post contains the explanation for why the misinformation is wrong or why originators of misinformation disseminated the false information.	"0"-no, "1"-yes, for the posts including multi-corrections, it is annotated as the number of corrections providing explanation divided by the total number of corrections mentioned in the post.
Graphic explanation	Whether contains the explanation in graphic form	"0"-no, "1"-yes, for the posts including multi-corrections, it is annotated as the number of corrections providing graphic explanation divided by the total number of corrections mentioned in the post.
Textual warnings of misinformation	Whether contains textual warnings before first mentioning the misinformation	"0"-no, "1"-yes, for the posts including multi-corrections, it is annotated as the number of corrections containing textual warnings divided by the total number of corrections mentioned in the post.
Graphic warnings of misinformation	Whether contains warnings in graphic form before first mentioning the misinformation	"0"-no, "1"-yes, for the posts including multi-corrections, it is annotated as the number of corrections containing graphic warnings divided by the total number of corrections mentioned in the post.
Influence of publisher	The influence of the publisher of the post	Be calculated as the sum of follower counts and other influence statistics, which is presented in Eq. (1).
Category	Topics of the post	Be determined by the topic of the misinformation being corrected, for the posts including multi-corrections, it is decided by the majority

Table 3
Parameters of models.

Algorithm	Parameter
SVR	kernal='rbf',C=0.6
KNN	n_neighbors=15, p=1, weights='distance'
Random Forest	n_estimators=330,max_depth=10,min_samples_leaf=1,max_features=3
XGBoost	learning_rate=0.05,n_estimators= 68,reg_alpha= 0.1, reg_lambda=0.9, gamma=0, subsample=0.75, colsample_bytree= 0.85,max_depth= 15, min_child_weight= 7

words, it means the number of users involved in social interactions with a post can represent the believability of the public to it. Besides, the effectiveness of correction means how many people can be persuaded to refute the misperception and to believe the truth. In this way, using the weighted sum of the number of likes L_j, retweets T_j and positive comments S_j can be a valid way to model the effectiveness of correction post j, E_j. Based on the various importance of social behaviors to believability in Kim and Dennis (2018), we take $\alpha = 0.4, \beta = 0.3, \gamma = 0.3$.

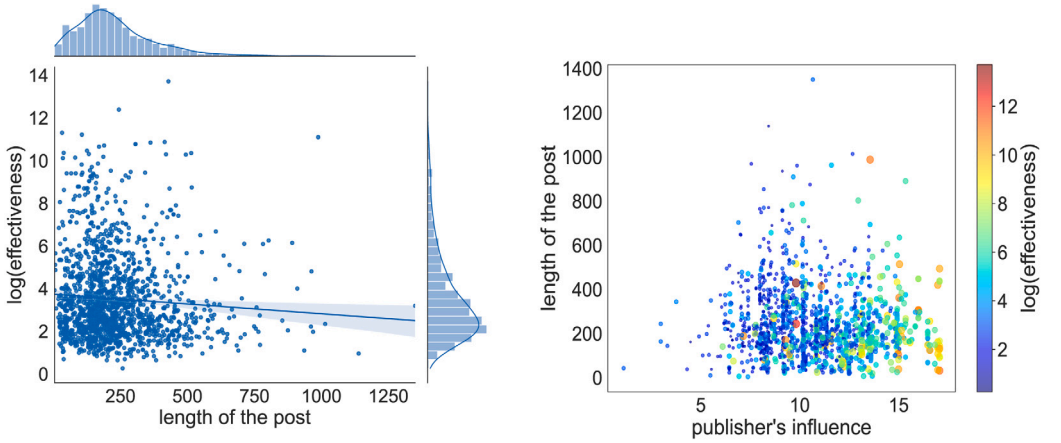
$$E_j = \alpha * L_j + \beta * T_j + \gamma * S_j \tag{2}$$

The sentiment analysis on comments is executed by a Python library named SnowNLP. It can handle Chinese text conveniently, including text segmentation, part-of-speech tagging, sentiment classification, etc.

4.3. Regression modeling

The examination of assumptions provides good understandings of effective mechanisms of misinformation correction on social media. However, it still cannot provide qualitative measurements of correction effectiveness. To solve this problem, we develop a data-driven prediction model with respect to the basic features of correction posts. We conduct our prediction in terms of typical machine learning methods, i.e., SVR, KNN, Random Forest, XGBoost. The dataset consists of 1487 correction posts. We randomly split it into the training set and testing set with a ratio of 7:3. The 5-fold validation is executed in the training set to select the model with the optimal parameters. The metrics used for evaluating the effectiveness of these models are Mean Absolute Error (referred to as MAE) and Coefficient of Multiple Determination (referred to as R^2). MAE represents deviations from the predicted values of regression models to annotations, and R^2 evaluates the explainability of independent variables to dependent variables in a regression model. Each regressor has been fine-tuned to the optimal parameters shown in Table 3. All models are built by a Python library named sklearn.

Subsequently, we make a thorough analysis of the dataset. We also apply the Spearman Correlation Analysis to verify the associations between the factors and the correction effectiveness. Moreover, the regression model is built to predict the efficacy of corrections. Results are displayed and discussed in Section 5.



(a) The effect of "length of the post"

(b) The effects of "influence of publisher" and "length of the post"

Fig. 2. Effects of "length of the post" and the "influence of publisher" on correction effectiveness.

5. Evaluation and discussion

In this section, the comprehensive analysis on COVID-related corrections and the experiments of the regression model are presented and discussed.

5.1. Data analysis results

This part demonstrates the effective determinants governing the correction efficacy and examines the veracity of the hypothesis. Subsequently, we verify the effects of factors on various topics of corrections. Finally, the influence of users involved in the dissemination of corrections is further explored.

Effective influencing factors The results of Spearman Correlation Analysis are shown in Table 4. The details and alternative explanations are discussed below.

Mentioning excessive original misinformation in corrections would not undermine people's believability within a short period after reading. In Table 4, the relation between this factor and correction effectiveness is not statistically significant, suggesting that the aforementioned hypothesis H1 is not supported. It reveals that excessive misinformation in a correction post would not reinforce the misperception of the public. Since our dataset is based on instant reactions from people, we speculate based on results that people can temporally recognize the misinformation clearly after reading corrections. Therefore, excessive original misinformation in correction posts cannot interfere with people's instant judgments on the veracity of information.

Concise corrections are more effective, which is recommended to be less than 500 words. Table 4 shows that the post length has a negative relation to correction effectiveness, which supports hypothesis H2. Fig. 2(a) visualizes the negative relationship between the "length of the post" and "correction effectiveness" on a log scale. (The right histograms and the top one represent the distribution of effectiveness on a log scale and the distribution on length of the post, respectively.) Moreover, it is suggested that the length of posts of current corrections is mainly distributed in [0,400] words. It should be mentioned that the correction effectiveness drops sharply when the length exceeds 500. Therefore, the length of posts is recommended to be less than 500 words.

The textual explanation seems to fail to improve correction. H3 argued that explanations for why misinformation is wrong or how it started to spread could correct misperceptions of people, especially graphic ones. Results from Table 4 suggest that the explanation does not significantly affect correction effectiveness and that the graphic explanation has a small positive effect on it, so H3 is half true. The finding is against common sense that alternative explanation makes corrections more feasible for people. Moreover, it should be stressed that graphic explanation is underestimated and rarely used (see Fig. 3), and most graphic explanations are of poor quality, e.g., screenshots of other correction posts, web pages, or chat. Concise and persuasive graphical explanations are appealed to be valued. Warnings in a tough tone make correction worse. Surprisingly, it is investigated that the textual and graphic warnings of misinformation both have negative associations with correction effectiveness. This observation is opposite to hypothesis H4. However, it is commonly believed that warnings of false information can make people more aware of a bias and think critically. Nowadays, mainstream media usually prefers to use eye-catching warnings to attract people's attention in correction posts (see Fig. 3), such as "xxx is wrong!", "xxx is not real!". These warnings are in demanding narrative and contain strong negative emotions, so it is easier to trigger Disconfirmation Bias (Nyhan & Reifler, 2010) which contributes to the backfire effect of corrections. Disconfirmation Bias means when people meet the arguments that challenge their worldview, their pre-beliefs will not change and even strengthen to argue against the opposing arguments. Hence, corrections should be demonstrated more gently and persuasively to avoid this backfire effect, not in a tough and strong tone.

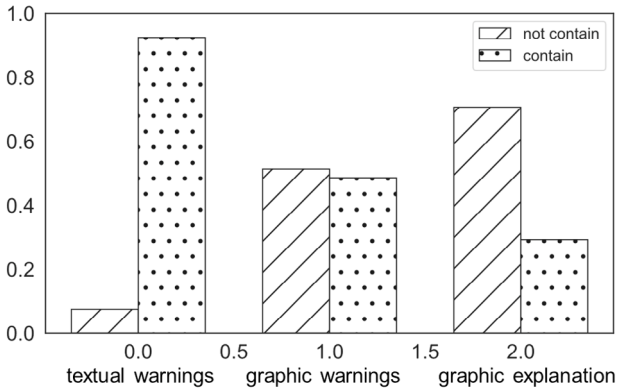


Fig. 3. Distributions of warnings and explanations in corrections.

Table 4
Spearman coefficient between potential factors and correction effectiveness.

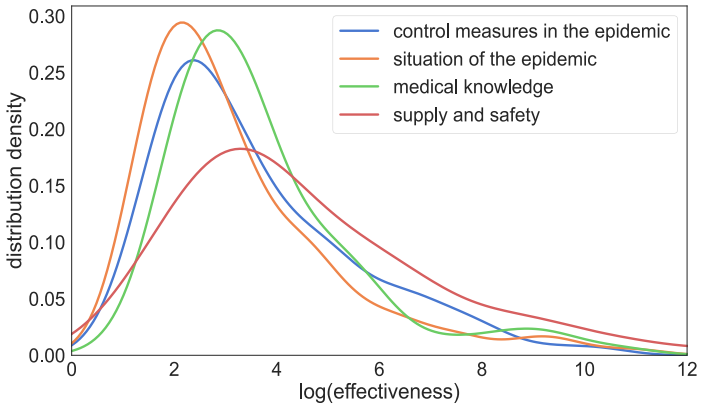
Factor	Coefficients	Sig.
Proportion of original misinformation	−0.033	0.208
Length of the post	−0.085**	0.001
Explanation	0.003	0.895
Graphic explanation	0.063*	0.016
Textual warnings of misinformation	−0.103**	0.000
Graphic warnings of misinformation	−0.104**	0.000
Influence of publisher	0.484**	0.000

The influence of publishers improves correction significantly. Finally, as H5 assumed, the publisher's influence positively and significantly relates to the correction effectiveness. Fig. 2(b) presents the combined effects of the "influence of publisher" and "length of the post" on the efficacy of correction. (Color and the size of data points represent correction effectiveness on a log scale.) It demonstrates that, as influence increases, the effectiveness is distributed in the range of more effective areas. Therefore, influential media should take more responsibility to combat misinformation.

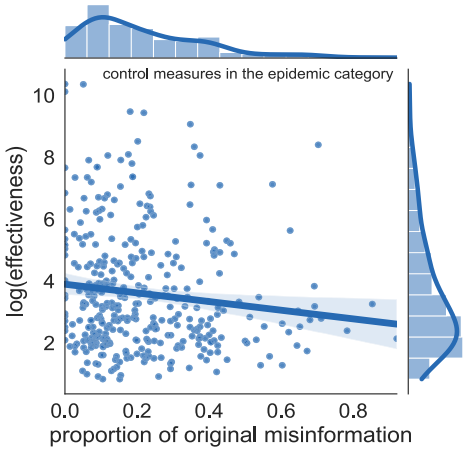
Influencing factors of correction effectiveness on different themes In Section 4.1, we classified four specific types of corrections based on the misinformation topics, namely, "control measures in the epidemic", "situation of the epidemic", "medical knowledge", and "supply and safety". Fig. 4(a) presents the distribution of correction effectiveness in all topics respectively. (Color represents different topics.) It is revealed that the corrections in supply and safety are the most effective, and the distributions of validity in other types are similar.

Subsequently, the effects of factors on each type were verified by Correlation Analysis. Results shown in Table 5 suggest that the effects of factors vary in different themes of corrections. For "situation of the epidemic", affective factors are the same as analysis results on overall corrections (referred to "overall result" below). However, for the "medical knowledge" type, the only influential factor is the influence of the publisher (0.391**, 0.000). The reason for this may be that corrections in "medical knowledge" contain more complex knowledge and are more difficult to understand than those in other themes, so the influencing factors in this type could be more complicated and need to be explored further. Finally, for the "control measures in the epidemic" type and "supply and safety" type, the textual warnings of misinformation and influence of publisher have effects on them just like overall results, but the length of posts, graphic warnings and graphic explanation have no influence. And surprisingly, the proportion of original misinformation has negative and significant effects on these two types (see Figs. 4(b) and 4(c), in each figure the right histograms and the top one represent the distribution of effectiveness on a log scale and the proportion of original misinformation, respectively.). Therefore, while correcting misinformation, various topics of events need different optimal formats of corrections.

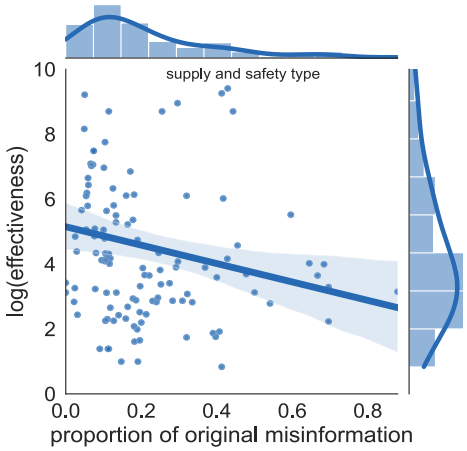
Involved users in the dissemination of corrections Motivated by RQ1, an analysis of users involved in the dissemination of corrections was conducted to identify their impacts on the efficacy of corrections. The posts of corrections were equally divided into two parts ranked by the effectiveness of the correction. Part A is more effective and the other part is referred to as Part B. The distribution of social interactions (i.e., "retweet", "support", which means leave supportive comments under the post, "criticize", which is contrary to "support") is presented in Fig. 5(a). Obviously, the number of overall users involved in Part A is approximately 13 times as one in Part B, which can be one of the reasons that Part A is more valid. Figs. 5(b) and 5(c) illustrate the distributions of users' followers in retweeting and supporting, respectively. It is indicated that social media influencers participate much more in Part A, which can be another reason for the better effectiveness. It is also observed that influencers retweet more rather than leave comments. Fig. 5(a) shows that supporting opinions and critical comments are almost equally distributed in both parts. Therefore, if influencers also leave supportive comments to guide public opinions, the efficacy of corrections can be improved.



(a) Distribution of correction effectiveness in various topics



(b) Effect of a factor in "control measures in the epidemic"



(c) Effect of a factor in "supply and safety"

Fig. 4. Correction effectiveness and influencing factor in different topics of corrections.

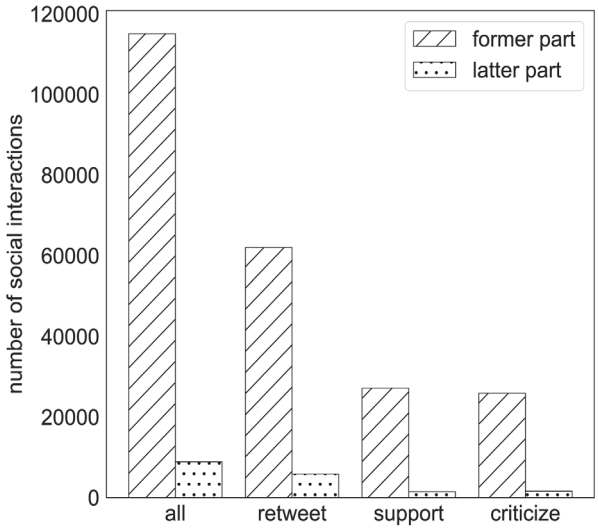
Table 5
Spearman coefficient between influencing factors and correction effectiveness in various topics of corrections.

Factor	Control measures in the epidemic		Situation of the epidemic		Supply and safety	
	Coefficients	Sig.	Coefficients	Sig.	Coefficients	Sig.
Proportion of original misinformation	−0.125*	0.019	0.012	0.737	−0.288**	0.001
Length of the post	0.004	0.939	−0.100**	0.006	0.021	0.816
Explanation	0.023	0.662	−0.013	0.725	0.120	0.193
Graphic explanation	−0.006	0.911	0.126**	0.001	0.049	0.592
Textual warnings of misinformation	−0.194**	0.000	−0.091*	0.012	−0.290**	0.001
Graphic warnings of misinformation	−0.065	0.218	−0.121**	0.001	−0.125	0.174
Influence of publisher	0.540**	0.000	0.451**	0.000	0.458*	0.000

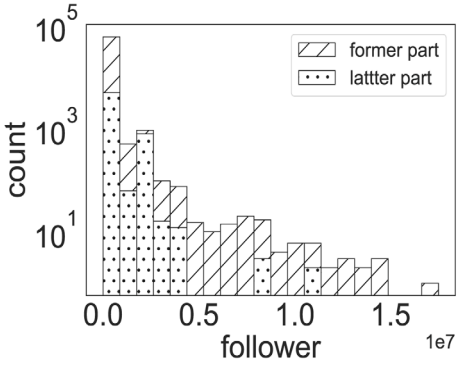
5.2. Regression modeling

This subsection illustrates the experiment results of the regression model and validates the sufficiency of the model through feature importance.

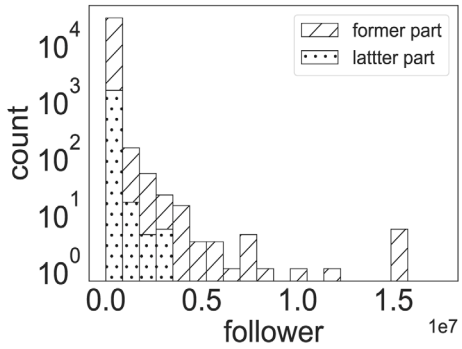
Based on the above analysis, the category of corrections has an association with their effectiveness. Accordingly, the input of the model contains the category and valid factors examined in Section 5.1 (i.e., length of the post, graphic explanation, textual warning



(a) Distribution on social interactions in two parts



(b) Distribution on followers of retweet users in two parts



(c) Distribution on followers of supportive users in two parts

Fig. 5. Distributions on social interactions and followers of users involved in the dissemination of the correction post.

Table 6
Evaluations of trained regression models.

Algorithm	Training set		Testing set	
	MAE	R^2	MAE	R^2
SVR	0.0867	0.3127	0.0849	0.3690
KNN	0.0005	0.9978	0.0793	0.4035
Random Forest	0.0639	0.6277	0.0784	0.4585
XGBoost	**0.0611**	**0.6611**	**0.0794**	**0.4638**

of misinformation, graphic warning of misinformation, and influence of publisher). During data processing, we apply a Leave-One-Out encoder to encode the category variable, which can avoid feature sparsity caused by One-hot encoding. Furthermore, other variables are normalized. We implemented regression models to predict the efficacy of corrections, as shown in Section 4.3. The regression modeling results are presented in Table 6, and each regressor has been tuned to the optimal parameters.

As Table 6 shows, SVR is underfitting and KNN is overfitting. The performance of XGBoost and Random Forest are comparable, whose MAE and R^2 are far better both in the training and testing set. We further compare the feature importance of these two outperformed regression models, which is shown in Table 7. In XGBoost, it is clearly seen that the influence of publisher, textual warnings of misinformation and category of misinformation have significant impacts on the prediction of correction effectiveness, and that the length of the post and graphic explanation have relatively smaller contributions, which captures the majority of associations found in the subsection "Effective influencing factor" of Section 5. The feature importance of Random Forest reflects too

Table 7
Feature importance of regression model.

Factor	Random Forest	XGBoost
	Feature importance	
Category	0.0895	0.1216
Length of the post	0.1982	0.0942
Graphic explanation	0.0212	0.0964
Textual warnings of misinformation	0.0639	0.3263
Graphic warnings of misinformation	0.0242	0.0843
Influence of publisher	0.6031	0.2772

much reliance on "influence of publisher", which hinders the model to learn the actual relationships among variables. Obviously, the XGBoost outperforms other regression models and learns the actual relation between the factors and correction effectiveness. Despite the good performance of XGBoost, the precision and robustness of the model can be further improved with a more large-scale dataset. we address this in Section 7.

6. Implications

This study sheds light on improving correction effectiveness on social media in several ways. Firstly, this study conducts manual labeling and evaluations on 1487 COVID-19 related corrections based on seven features. It provides a dataset for researchers interested in mechanisms of misinformation correction on social media, which fills the gap in the field. Specifically, researchers can use this dataset to train advanced machine learning models to automatically predict correction effectiveness. Also, the features within can be expanded to further explore determinants of correction effectiveness, according to needs.

Secondly, the findings of this study provide validations for related theories, demonstrate effective factors associated with correction effectiveness on current social media, and offer insightful and practical suggestions. Namely, compared to the previous studies, this study provides evidence for overkill effect (Ecker et al., 2019; Lewandowsky et al., 2012; Schwarz et al., 2007), and reminds researchers to take a second think of the applicability of familiar backfire effect (Ecker et al., 2017, 2020; Lewandowsky et al., 2012) and mental model of misinformation (Ecker et al., 2011, 2010) on current social media. Besides, it reminds content publishers to notice the intrinsically psychological factors interfering with the process of adoption by people, instead of paying much attention to catching eyes, e.g., the usage of strong and explicit warning. Also, the social platform can utilize these findings to promote the credibility of corrections published by influential users. according to findings, the publishers' influence has a significant impact on correction effectiveness. Therefore, it could help if the social media platform emphasizes the influence of publishers with ratings.

Thirdly, this study proposes a regression model to predict correction effectiveness, with the input of basic features of corrections. This can guide practitioners to revise their corrections before publishing, leading to ideal efficiency.

7. Limitations and future work

First, we evaluate correction effectiveness based on social behaviors, such as retweeting, liking, commenting, which are related to people's instant reactions to corrections. But the continued effects of misinformation can be persistent. Therefore, it is important to verify the effectiveness of corrections over a long period of time. For example, we can track users' positions on an event through their posts, browsed content, posts they interact with, etc. However, these tasks might be challenging due to the availability of data. Second, the effective factors associated with the efficacy of corrections may change, while considering the effectiveness in a temporal sequence. In addition, because our work relies on 1487 correction posts, the dataset can be expanded to acquire more universal conclusions, and to construct a high-precision and robust model.

8. Conclusions

Based on an analysis of the COVID-19 related dataset, we demonstrate five findings on the good practice of effective corrections. First, mentioning excessive original misinformation in corrections would not undermine people's believability within a short period after reading. Second, concise corrections are more effective, which is recommended to be less than 500 words. Third, persuasive graphic explanations are appealed to be valued. Fourth, corrections should be demonstrated more gently and persuasively, not in a tough and strong tone. Fifth, Influential media should take more responsibility to combat misinformation. In summary, we appeal for the corrections which are short, concise, persuasive, rich in graphics, in a gentle tone, and published by influential users. Also, we reveal that the effects of the influencing factors on corrections vary among various topics. Therefore, it should be considered while debunking. At the end of this work, we build the regression model to predict the effectiveness of corrections, which is proven that the model learns the actual association among the basic features of correction and its effectiveness. This model can guide practitioners to revise their corrections before publishing and achieve effective corrections.

CRediT authorship contribution statement

Yuqi Zhang: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Bin Guo:** Conceptualization, Supervision, Writing – review & editing. **Yasan Ding:** Writing – review & editing, Methodology. **Jiaqi Liu:** Writing – review & editing. **Chen Qiu:** Writing – review & editing. **Sicong Liu:** Writing – review & editing. **Zhiwen Yu:** 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.

Acknowledgments

This work was partially supported by the National Science Fund for Distinguished Young Scholars (62025205), National Key R&D Program of China (2019QY0600), the National Natural Science Foundation of China (No. 61960206008, 61725205, 62102317,62002292), and Natural Science Basic Research Plan in Shaanxi Province of China (2020JQ-207).

References

Alda, A., Bass, E. R., Chedd, G., Constantinou, C., O'Connell, C., Schneider, H., et al. (2012). *The debunking handbook*. Stony Brook University. School of Journalism.

Andrea, E., & Radvansky, G. A. (2020). Failure to accept retractions: A contribution to the continued influence effect. *Memory & Cognition*, *48*(1), 127–144.

Bautista, J. R., Zhang, Y., & Gwizdka, J. (2021). Healthcare professionals' acts of correcting health misinformation on social media. *International Journal of Medical Informatics*, *148*, Article 104375.

Bode, L., & Vraga, E. K. (2018). See something, say something: correction of global health misinformation on social media. *Health Communication*, *33*(9), 1131–1140.

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., et al. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, *1*(3).

Budak, C., Agrawal, D., & El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on world wide web* (pp. 665–674).

Burel, G., Farrell, T., & Alani, H. (2021). Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter. *Information Processing & Management*, *58*(6), Article 102732.

Chen, S., Xiao, L., & Mao, J. (2021). Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management*, *58*(5), Article 102665.

Dai, Y., Yu, W., & Shen, F. (2021). The effects of message order and debiasing information in misinformation correction. *International Journal of Communication*, *15*, 21.

Desai, S. C., & Reimers, S. (2019). Comparing the use of open and closed questions for web-based measures of the continued-influence effect. *Behavior Research Methods*, *51*(3), 1426–1440.

Ecker, U. K., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, *6*(2), 185–192.

Ecker, U. K., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane!—No, actually it was a technical fault: Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology*, *64*(2), 283–310.

Ecker, U. K., Lewandowsky, S., Jayawardana, K., & Mladenovic, A. (2019). Refutations of equivocal claims: No evidence for an ironic effect of counterargument number. *Journal of Applied Research in Memory and Cognition*, *8*(1), 98–107.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100.

Ecker, U. K., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*, *111*(1), 36–54.

Fuyong, Y., Jing, F., Qianqian, F., & Xufeng, C. (2012). A method to reduce the impact of zombie fans in micro-blog. *Data Analysis and Knowledge Discovery*, *28*(5), 70–75.

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2020). The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys*, *53*(4), 1–36.

He, X., Song, G., Chen, W., & Jiang, Q. (2012). Influence blocking maximization in social networks under the competitive linear threshold model. In *Proceedings of the 2012 siam international conference on data mining* (pp. 463–474). SIAM.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*. Yale University Press.

Jerit, J., & Barabas, J. (2012). Partisan perceptual bias and the information environment. *The Journal of Politics*, *74*(3), 672–684.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420.

Kim, A., & Dennis, A. (2018). Says who?: How news presentation format influences perceived believability and the engagement level of social media users. In *Proceedings of the 51st hawaii international conference on system sciences*.

Lederer, E. (2020). *UN chief antonio guterres: misinformation about COVID-19 is the new enemy*. NY: TIME New York.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Luo, J., Xue, R., & Hu, J. (2020). COVID-19 infodemic on Chinese social media: A 4P framework, selective review and research directions. *Measurement and Control*, *53*(9–10), 2070–2079.

McCracken, G. (1989). Who is the celebrity endorser? Cultural foundations of the endorsement process. *Journal of Consumer Research*, *16*(3), 310–321.

Nguyen, N. P., Yan, G., & Thai, M. T. (2013). Analysis of misinformation containment in online social networks. *Computer Networks*, *57*(10), 2133–2146.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Pian, W., Chi, J., & Ma, F. (2021). The causes, impacts and countermeasures of COVID-19 “infodemic”: A systematic review using narrative synthesis. *Information Processing & Management*, *58*(6), Article 102713.

Rich, P. R., & Zaragoza, M. S. (2020). Correcting misinformation in news stories: An investigation of correction timing and correction durability. *Journal of Applied Research in Memory and Cognition*, *9*(3), 310–322.

Saxena, A., Hsu, W., Lee, M. L., Leong Chieu, H., Ng, L., & Teow, L. N. (2020). Mitigating misinformation in online social network with top-k debunkers and evolving user opinions. In *Companion proceedings of the web conference 2020* (pp. 363–370).

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, *39*, 127–161.

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? In *Psychology of learning and motivation. Vol. 41* (pp. 265–292). Elsevier.

Shu, K., Bernard, H. R., & Liu, H. (2019). Studying fake news via network analysis: detection and mitigation. In *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 43–65). Springer.

Song, C., Hsu, W., & Lee, M. L. (2017). Temporal influence blocking: Minimizing the effect of misinformation in social networks. In *2017 IEEE 33rd international conference on data engineering* (pp. 847–858). IEEE.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460–480.

Tong, G., & Wu, W. (2018). On misinformation containment in online social networks. In *32nd Conference on neural information processing systems*.

Van den Broek, P., Young, M., Tzeng, Y., Linderholm, T., et al. (1999). The landscape model of reading: Inferences and the online construction of a memory representation. In *The construction of mental representations during reading* (pp. 71–98).

Vraga, E. K., & Bode, L. (2020). Correction as a solution for health misinformation on social media. *American Journal of Public Health*, *110*(S3), S278–S280.

Wilkes, A., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology*, *40*(2), 361–387.

Wu, P., & Pan, L. (2017). Scalable influence blocking maximization in social networks under competitive independent cascade models. *Computer Networks*, *123*, 38–50.

Yang, L., Li, Z., & Giua, A. (2020). Containment of rumor spread in complex social networks. *Information Sciences*, *506*, 113–130.

Yang, D., Liao, X., Shen, H., Cheng, X., & Chen, G. (2018). Dynamic node immunization for restraint of harmful information diffusion in social networks. *Physica A: Statistical Mechanics and its Applications*, *503*, 640–649.

Zareie, A., & Sakellariou, R. (2021). Minimizing the spread of misinformation in online social networks: A survey. *Journal of Network and Computer Applications*, Article 103094.

Zhang, Y., Yang, W., & Du, D.-Z. (2021). Rumor correction maximization problem in social networks. *Theoretical Computer Science*, *861*, 102–116.

Zhou, C., Xiu, H., Wang, Y., & Yu, X. (2021). Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on COVID-19. *Information Processing & Management*, *58*(4), Article 102554.

Zrnec, A., Poženel, M., & Lavbič, D. (2022). Users’ ability to perceive misinformation: An information quality assessment approach. *Information Processing & Management*, *59*(1), Article 102739.