

Original Paper

Health-Related Rumor Debunking on Sina Weibo in China (2009-2024): 15-Year Retrospective Infodemiology Study

Yuan Fang, BMgmt; Chengwu He; Yejinxuan Hu, BMgmt; Xianyun Tian, PhD

College of Management Science, Chengdu University of Technology, Chengdu, Sichuan, China

Corresponding Author:

Xianyun Tian, PhD
College of Management Science
Chengdu University of Technology
No. 1, East Third Road, Erxianqiao, Chenghua District
Chengdu, Sichuan 610059
China
Phone: 86 17794550089
Email: xianyuntian@cdut.edu.cn

Abstract

Background: Understanding how health-related rumor debunking evolves and spreads on social media is critical for public health communication and policy. Existing research, however, has been largely crisis-centered—dominated by studies of specific events, such as the COVID-19 pandemic—and offers limited insight into the longer-term patterns of thematic evolution, demographic targeting, and engagement dynamics of official debunking practices.

Objective: This study aimed to provide an integrated understanding of official health rumor debunking in China over a 15-year period by delineating its thematic evolution, demographic disparities, and features associated with engagement and information diffusion.

Methods: We collected rumor debunking posts published on Sina Weibo between 2009 and 2024. A 2-stage classification pipeline using Sentence Transformer Fine-Tuning was constructed to identify health-related debunking posts. We applied BERTopic (BERT: Bidirectional Encoder Representations from Transformers) to map the thematic landscape and used a large language model to extract demographic mentions (gender, age, and social roles), and health content referenced in the posts. Finally, we used Extreme Gradient Boosting regression models with Shapley Additive Explanations to quantify the relative contributions of predictors to engagement and information diffusion, incorporating a multidimensional feature space spanning user metadata, content characteristics, and temporal and contextual attributes.

Results: Across 377,520 rumor debunking posts from 24,742 blue-verified accounts, 93,799 posts (24.8%) were health-related. Health debunking volume surged during the COVID-19 pandemic and remained elevated above prepandemic levels through 2024, although its relative share declined markedly in 2024, indicating a shift away from pandemic-centered topics, even as overall debunking activity continued to rise. Prevalent themes included vaccines, debunking reports, personal care, and sleep hygiene. Temporal trajectories fell into 3 patterns: event-driven spikes, sustained growth, and recurrent fluctuations. Demographic mentions were uneven: women were referenced more often than men, youth were the most frequently mentioned age group, and older adults were the least mentioned. Engagement was predicted primarily by source-level reach: follower count was the strongest predictor of reposts, comments, and likes alike, showing a nonlinear, threshold-like pattern in each case, whereas the secondary predictors differed across the 3 interaction types.

Conclusions: This 15-year longitudinal analysis shows that health-related debunking on Sina Weibo was strongly shaped by major public health crises, surging during the COVID-19 pandemic and remaining elevated, while broadening toward routine, lifestyle-related concerns. Across all interaction types, engagement was shaped more by source reach than by the content of corrections. Although most posts addressed the general public, those with explicit demographic references revealed uneven representation across gender and age groups, with distinct health concerns linked to each. Taken together, these findings reveal a structural asymmetry in the debunking ecosystem: content is diversifying while distribution remains governed by source-level reach, suggesting that platform-level mechanisms that help credible but less-followed sources circulate corrections may be valuable for amplifying reliable health information.

Keywords: infodemiology; health misinformation; rumor debunking; social media; health communication; public health surveillance

Introduction

The rapid expansion of the internet has lowered barriers to information dissemination, enabling content to circulate through online networks at unprecedented speed [1,2]. In such environments, rumors, unverified claims that spread widely in the absence of conclusive evidence, can proliferate, distort public opinion, amplify anxiety, and, in some cases, lead to serious real-world consequences [3-7].

Among the many forms of online misinformation, health-related rumors are particularly concerning because they involve issues closely tied to individual safety and public well-being [8]. At the same time, the specialized nature of medical knowledge often makes it difficult for nonexperts to assess information credibility, rendering many users vulnerable to misleading claims [9-13]. When such misleading claims take root among the public, they can disrupt public health decision-making, erode institutional trust, and produce severe real-world consequences [14,15]. During the COVID-19 pandemic, for example, false claims that consuming methanol or alcohol-based disinfectants could eliminate the virus resulted in hundreds of preventable deaths and thousands of hospitalizations worldwide [16].

To mitigate these risks, governments, public health agencies, social media platforms, and news organizations have increasingly adopted rumor debunking strategies to curb the spread of misleading health information and restore public trust [17-19]. On social media platforms, official debunking posts play an increasingly important role in countering misinformation by providing corrective information from authoritative sources and helping to mitigate the adverse effects of false claims [20,21]. Rather than merely reacting to individual falsehoods, these debunking messages function as systematic institutional interventions embedded within broader information governance structures. As such corrective communication becomes increasingly central to public health, understanding how health-related debunking operates in real-world social media environments has emerged as an urgent priority for both researchers and practitioners.

Existing research on health misinformation and rumor debunking has advanced along several complementary lines. Prior studies have examined the thematic characteristics of health misinformation, identifying recurring issue domains such as public health crises, diseases, and diet- and nutrition-related claims [22,23]. Other work has investigated who is susceptible to health misinformation, highlighting the roles of demographic and psychological factors in shaping vulnerability to misleading information [12,24]. Researchers have also explored how misinformation spreads across social media platforms, documenting diffusion dynamics and network mechanisms [5,25]. In parallel, substantial attention has been devoted to misinformation governance, including automated identification and monitoring systems

[4,26], evaluations of corrective and debunking strategies [17,18], and the mechanisms and temporal durability of prebunking (inoculation) interventions [19,27]. Finally, researchers have examined audience responses to corrections, identifying factors that influence the acceptance of or resistance to debunking messages [18,28,29]. Despite this growing body of literature, several important gaps remain.

First, existing studies on health-related debunking have largely focused on specific public health crises—most notably COVID-19—thereby providing episodic snapshots rather than a longitudinal understanding of evolving discourse [15]. For example, Yang et al [30] examined rumor characteristics during the COVID-19 pandemic, whereas Sicilia et al [31] focused on the Zika virus. Although such crisis-centered approaches provide valuable insights into debunking dynamics under exceptional conditions, they are typically confined to relatively short time windows. Consequently, existing studies still lack a macro-level understanding of how health debunking themes are structured and how they evolve over time on social media platforms. Developing such a longitudinal perspective is important for informing the strategic allocation of debunking resources and supporting long-term public health governance planning [13,32,33].

Second, prior research has carefully examined who is susceptible to health misinformation. Li and Shu [34], for instance, identified demographic attributes such as gender and age as key determinants of vulnerability. Beyond demographics, cognitive traits have also been given substantial attention: Chua and Banerjee [35] highlighted the vulnerability of “naïve believers” who view knowledge as static, while Pennycook and Rand [36] suggested that susceptibility is often associated with cognitive miserliness, particularly among individuals who rely on intuition rather than analytic reasoning. However, much of this evidence is derived from self-reported surveys or small-scale experiments. As a result, existing findings may lack ecological validity and may not fully capture which demographic groups are explicitly referenced in real-world debunking narratives or what kinds of health concerns are associated with them within platform-based information ecosystems [37]. In practice, understanding which groups are explicitly mentioned as targets or affected audiences and what health risks are associated with these groups is crucial for precision debunking, improving intervention effectiveness, and optimizing resource allocation.

Third, existing research on the diffusion of health debunking information has primarily focused on differences across debunking content types, determinants of user engagement, and variations in debunking strategies. For example, Yang et al [30] examined the effects of rumor categories (dread vs wish) and rhetorical strategies, while Chua and Banerjee [35] investigated how different formats (text vs image) influence sharing intentions depending on user epistemic beliefs. Methodologically, however,

this line of work predominantly relies on linear statistical models, such as regression analysis. While useful for estimating average effects, these approaches are often ill-suited for the skewed and nonlinear nature of social media diffusion data. As Sicilia et al [31] noted, online information diffusion is shaped by complex network structures and long-tail distributions of influence (eg, follower counts), conditions that frequently violate the assumptions of traditional regression models. Consequently, how multidimensional factors—spanning source, content, and contextual attributes—jointly predict different modes of public engagement remains insufficiently understood. Uncovering these nonlinear mechanisms is crucial for platforms to optimize resource allocation and enhance the visibility of corrective communication.

Against this backdrop, this study conducted a systematic analysis of health-related debunking posts published by verified accounts on Sina Weibo (Beijing Weimeng Chuangke Network Technology Co., Ltd.), covering the period from August 2009 to December 2024. Notably, although these 3 research gaps have each attracted independent scholarly attention, existing studies have rarely examined them in conjunction. Yet the 3 dimensions correspond to successive stages of the same corrective communication process, from what is produced, to whom it addresses, to how far it travels, and the overall functioning of the debunking ecosystem may depend on their interaction rather than on any single dimension alone. Through this large-scale longitudinal investigation, we therefore aimed to advance an integrated, macro-level understanding of official health debunking practices in China by addressing the following research questions (RQs):

- RQ1: What characterizes the thematic landscape of official health debunking in China, and how have these themes evolved over the past 15 years?

- RQ2: Who are the targeted and affected groups mentioned in official health rumor debunking posts, and how do the associated health risks and concerns differ across these groups?
- RQ3: What are the primary predictors of public engagement with official health debunking information, and how do their marginal contributions differ across interaction modes such as reposts, comments, and likes?

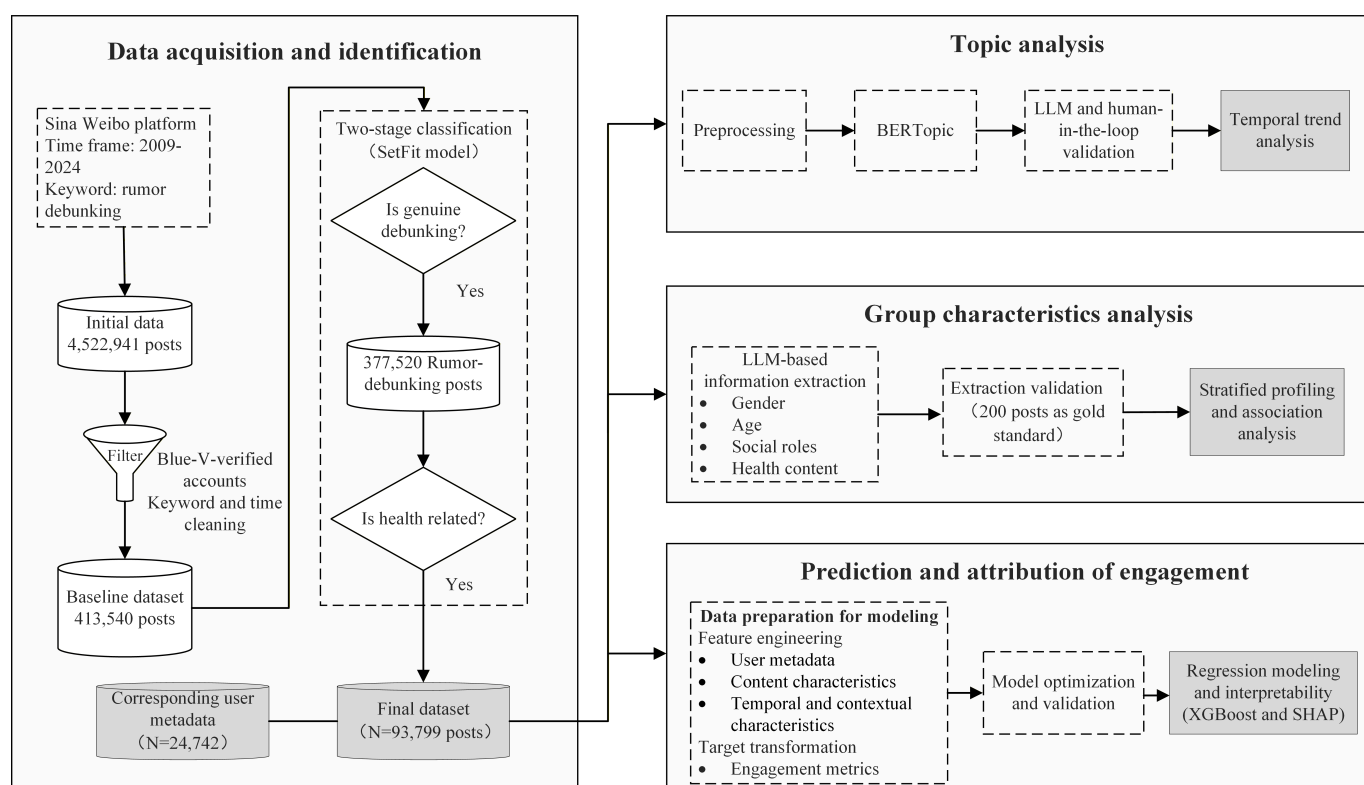
By addressing these questions, this study not only offers a comprehensive account of China's official health debunking efforts within a rapidly evolving social media environment but also provides actionable insights for public health authorities and platform administrators seeking to strengthen precise and efficient governance of health misinformation.

Methods

Study Design

This study adopted a retrospective infodemiology design to examine health-related rumor debunking posts on Sina Weibo over a period of 15 years (2009-2024). Our analytical framework integrates natural language processing, large language models (LLMs), and interpretable machine learning to address three core dimensions: (1) the thematic evolution of health-related debunking, (2) the demographic profiles of the groups targeted or affected in debunking narratives, and (3) the key predictors associated with information diffusion and user engagement. The overall research workflow is summarized in [Figure 1](#).

Figure 1. Schematic of the research framework. BERT: Bidirectional Encoder Representations from Transformers; LLM: large language modeling; SHAP: Shapley Additive Explanations; SetFit: Sentence Transformer Fine-Tuning.



Data Acquisition and Prefiltering

Sina Weibo is one of China's most influential social media platforms, reporting 587 million monthly active users [38]. Its large-scale and diverse user base make it a valuable source for examining public discourse related to health rumor debunking.

Using the Chinese keyword “rumor debunking” (辟谣), we retrieved posts published between August 1, 2009, and December 31, 2024. The initial corpus comprised 4,522,941 posts authored by 24,801 users.

To improve data quality, we applied a multistep filtering pipeline. First, based on Weibo's verification labels, we retained 493,956 posts published by “Blue V” (blue-verified) accounts, which are officially certified government agencies, media outlets, and organizational users. This restriction was motivated by prior evidence that official entities tend to provide more authoritative and effective rumor corrections [39,40]. Second, we cleaned the data by removing 80,416 posts that either did not contain the keyword or fell outside the specified time window, yielding 413,540 candidate posts. Third, we applied classifiers, as detailed in the “Classification Strategy and Model Selection” section, to identify genuine debunking posts and further categorize them into health-related versus other topics.

Subsequently, we retrieved user-level metadata using the user IDs associated with the identified genuine debunking posts. For each post, we compiled a comprehensive feature set. Post-level attributes included the text content, publication time, and engagement metrics (likes, comments, and reposts), whereas user-level metadata (when available)

included user ID, nickname, profile description, verification type, follower and following counts, and membership level. Account metadata for 948 users (1723 posts) were unavailable due to account anomalies (eg, self-deactivation or platform restrictions), but the corresponding posts were retained to preserve coverage for textual and thematic analyses. Ultimately, the final dataset contained 377,520 posts from 24,742 blue-verified accounts, including 93,799 health-related rumor debunking posts and 283,721 non-health-related rumor debunking posts. For the analysis of publisher verification types, posts with missing or invalid verification-type information were excluded, resulting in a subset of 282,039 posts. Publisher verification-type distributions between health-related and non-health-related posts were compared using a Pearson chi-square test, with Cramér *V* used to quantify the effect size.

Classification Strategy and Model Selection

Overview

Following standard practices in text classification [41], we adopted a 2-stage pipeline to improve task tractability and classification accuracy. Specifically, we first identified genuine rumor debunking posts from the full corpus and then classified the identified debunking posts as health related versus non-health related. This staged design reduces task complexity and enables models to learn more discriminative features, thereby improving overall performance [42,43]. To ensure the reliability of this pipeline, we relied on 2 core components: rigorous human annotation to establish high-

quality gold-standard labels and systematic model evaluation to select the most robust classifier.

Annotation Procedure and Label Definitions

We developed the annotation guidelines through an iterative process. The first author drafted an initial guideline after reviewing 500 randomly sampled posts. Next, the first and second authors completed multiple rounds of independent annotation, each round using a newly sampled set of 500 posts generated with different random seeds. The guideline was refined after each round until disagreements were largely eliminated and the criteria stabilized.

A post was labeled as debunking if it explicitly identified specific information as a rumor. Debunking posts were further labeled as health-related if they concerned human health, including (but not limited to) psychology, diet, medical treatments, disease prevention, lifestyle habits, infectious diseases, or public health events.

Using the finalized criteria, 3 authors participated in the final annotation. An additional 500 posts were sampled using a fixed random seed, and the first 2 authors independently annotated this set to ensure reproducibility. Interannotator reliability was assessed using Krippendorff α , which accommodates different data types and numbers of coders; values above 0.80 are commonly interpreted as strong reliability [44]. After α exceeded 0.80 for both classification tasks, any remaining disagreements were resolved through discussion, with the third author mediating when necessary, to produce the final gold-standard datasets. The resulting α values were 0.835 for debunking identification and 0.909 for health topic classification.

Model Evaluation and Selection

To select the best-performing classifier, we evaluated a range of representative approaches in natural language processing using 5-fold cross-validation on the gold-standard datasets. We tested traditional machine learning base-lines (support vector machine, random forest, Extreme Gradient Boosting [XGBoost], and logistic regression), a deep learning sequence model (attention-based bidirectional long short-term memory), and transformer-based pretrained language models (bert-base-chinese and Chinese-roberta-wwm-ext). We also evaluated Sentence Transformer Fine-Tuning (SetFit), a label-efficient framework that fine-tunes sentence transformers with relatively few labeled examples [45]. For SetFit, we compared multiple backbone models (bert-base-chinese, Chinese-bert-wwm-ext, paraphrase-multilingual-MiniLM-L12-v2, and Chinese-roberta-wwm-ext) to identify the most suitable option for our Chinese corpus, which was subsequently deployed to classify the full candidate dataset.

Topic Analysis

Topic Identification

To uncover latent themes in health-related rumor debunking posts, we applied BERTopic (BERT: Bidirectional Encoder Representations from Transformers), a topic

modeling framework that combines transformer-based sentence embeddings with density-based clustering [46]. Unlike probabilistic bag-of-words models such as Latent Dirichlet Allocation, which rely on document-topic distributions, BERTopic captures contextual semantic relationships at the sentence level, making it particularly suitable for short and colloquial Weibo texts [47].

Before topic modeling, we conducted light preprocessing to reduce platform-specific noise while preserving sentence structure for embedding generation. Specifically, we removed topic hashtags enclosed by “#” (eg, #health-related rumor-debunking#), user mentions starting with “@” (eg, @People’s Daily), and URLs. After cleaning, 2 posts with empty content were excluded.

BERTopic was then applied to the processed corpus. Texts were embedded using the sentence transformer model paraphrase-multilingual-mpnet-base-v2 [48]. We reduced the embedding dimensionality to 5 components using Uniform Manifold Approximation and Projection and clustered the reduced embeddings with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN uses a soft-clustering approach that accommodates structural noise by modeling ambiguous texts as outliers (topic –1) [46]. By preventing the forced assignment of unrelated documents, this mechanism improves topic coherence and semantic interpretability.

For each cluster, we extracted the 10 most representative terms using class-based Term Frequency–Inverse Document Frequency, which estimates term importance across document clusters rather than individual documents, thereby generating robust topic-word distributions [46]. To improve interpretability, we filtered topic keywords using a publicly available Chinese stop-word list (Baidu stop-word list) and removed numerical digits (0–9) to avoid noninformative tokens in topic descriptions.

Topic Validation and Labeling

To verify that the HDBSCAN outlier exclusion (topic –1) was temporally unbiased and did not distort the longitudinal dataset, we compared the month-by-month distribution of outlier versus retained posts across the 2009 to 2024 study window using Spearman rank correlation and the Kolmogorov-Smirnov test. Having confirmed the temporal robustness of the dataset, we then quantitatively evaluated the semantic quality of the retained clusters by calculating topic coherence and topic diversity scores based on the top representative keywords.

While these quantitative metrics confirmed the structural robustness of the clusters, assigning accurate labels to colloquial social media discourse requires deeper semantic comprehension. Therefore, we used a human-in-the-loop framework to finalize the topic representations. For each of the top 10 topics, we extracted the 10 highest-ranked class-based Term Frequency–Inverse Document Frequency keywords, along with the 5 most representative documents (selected based on the highest count of matched core keywords, prioritizing longer texts when match scores were

equal to ensure maximum contextual richness for the LLM). We then prompted Doubao (Doubao-Seed-2.0-pro-260215), a leading Chinese LLM according to the SuperCLUE [49] benchmark, to generate preliminary academic labels using the extracted keywords and documents as context. Finally, one of the authors reviewed each generated label against the raw representative documents and the broader cluster content; any labels deemed potentially inaccurate or ambiguous were discussed and refined through team consensus among the co-authors to ensure semantic accuracy.

Temporal Trend Analysis

To examine how health-related rumor debunking themes evolved over the study period, we assigned each post its corresponding BERTopic topic label and aligned these labels with publication time stamps. We then computed and visualized the topic frequency over time, which enabled us to identify shifts in prominence and the emergence of new themes across years.

Group Characteristics Analysis

Automated Information Extraction via LLMs

Given the finite resources available for large-scale information governance, accurately identifying vulnerable groups is critical for efficiently targeting health-related rumor debunking. To characterize the populations referenced in debunking posts and examine their associations with specific health topics, we used DeepSeek-V3.2, an LLM with strong Chinese language understanding and instruction-following capability [50], to build a standardized information extraction pipeline. The model was accessed via an OpenAI-compatible API and applied to the full dataset in batch processing mode. We designed an extraction schema comprising 4 key variables to capture the multidimensional nature of digital health vulnerability: gender and age provide the core demographic strata; social roles (eg, occupation) capture contextual vulnerability beyond biological attributes; and health content reflects the specific medical concerns linked to misinformation.

To transform diverse textual features into quantifiable demographic variables, we enforced strict standardization through a rule-embedded system prompt. We designed a comprehensive instruction set that directs the model to map extracted features based on predefined criteria. For gender, diverse referents ranging from specific terms such as “Mr.” to implicit relational terms such as “mother” are mapped to unified “male” or “female” categories. Regarding age, we established a hybrid classification framework anchoring definitions in legal and medical standards. Specifically, minors (aged <18 y) and the older adults (aged ≥60 y) follow the strict definitions of the Law on the Protection of Minors [51] and the Law on Protection of the Rights and Interests of the Elderly of the People’s Republic of China [52], respectively. The intermediate spectrum is further stratified into youth (aged 18–44 y) and middle-aged (aged 45–59 y), drawing upon epidemiological evidence regarding the onset of chronic noncommunicable diseases [53].

Furthermore, we addressed the structural complexity of social media content, particularly “listicle” style posts where a single text often debunks multiple unrelated rumors simultaneously. To resolve this, we integrated structural demonstrations (few-shot examples) into the prompt. This guided the model to parse complex texts and segment distinct semantic units, ensuring that social roles and specific health content were extracted as structured lists corresponding to each subtopic, thereby preserving semantic granularity and completeness.

Validation of the Extraction Framework

Before large-scale deployment, we validated the reliability of this extraction framework using a 200-post gold-standard set annotated independently by 2 coders. Interannotator agreement (IAA) was first calculated to establish the reliability of the human annotations. The model’s extraction performance was assessed by comparing the outputs of DeepSeek-V3.2 with the human-adjudicated ground truth. Evaluation metrics were selected to match the structural characteristics of each extracted attribute. For standardized categorical variables, specifically the mapped gender and age groups, Krippendorff α was used to measure IAA, and the Macro F_1 -score was used to evaluate model performance. For unordered entity lists, such as raw demographic mentions and identity labels, a set-based F_1 -score was applied. Finally, for free-text health narratives, BERTScore was used to assess semantic similarity beyond surface-level lexical overlap. The validation results demonstrated high IAA across all dimensions, and the pipeline achieved strong alignment with the adjudicated benchmark, thereby supporting its robustness for extracting group mentions and health content from the full corpus (n=93,799). Detailed validation metrics, including overall performance evaluations and category-specific metrics for gender and age variables, are provided in Tables S1–S3 in [Multimedia Appendix 1](#).

Demographic and Thematic Association Analysis

On the basis of the standardized outputs returned by the validated pipeline, we conducted descriptive and association analyses to characterize group mentions and their topical distributions. We first calculated the frequency of extracted gender and age groups across the top BERTopic topics. We then performed word frequency analysis on the extracted social roles and health content to summarize salient health narratives for each demographic subgroup using Chinese word segmentation and stop-word filtering. The most frequent keywords were computed to identify prominent social identities and compare differences in health concerns across demographic strata.

Prediction and Explanation of Information Diffusion via XGBoost and Shapley Additive Explanations

Overview

To examine the complex factors associated with the propagation of health-related rumor debunking posts, it was necessary

to account for the nonlinear dynamics and diminishing returns inherent in social media diffusion [54]. To capture these patterns while maintaining interpretability, we adopted an interpretable machine learning framework combining XGBoost Regressor for predicting continuous engagement outcomes [55], with SHAP (Shapley Additive Explanations) [56] for model interpretation. This framework enabled us to estimate the expected magnitude of user engagement while quantifying the marginal contributions of source characteristics, content attributes, and contextual cues, thereby facilitating a more transparent understanding of the mechanisms underlying rumor debunking diffusion [57].

Feature Engineering and Target Variable Transformation

To mitigate potential omitted-variable bias and capture the multidimensional context of information diffusion on social media platforms, we engineered features across 3 major dimensions: user metadata, content characteristics, and temporal and contextual attributes.

First, regarding user metadata, we included account verification type and membership category as categorical features, while ordinal membership ranks were label encoded. To reduce skewness and stabilize variance, highly skewed account-level metrics—including follower count, following count, and status count—were transformed using a base-10 logarithm.

Second, for content characteristics, we incorporated 4 primary dimensions: thematic categories, multimedia elements, text length, and sentiment polarity. Thematic categories were derived from BERTopic, and multimedia elements were represented by the presence of attached images and videos. Regarding text length, it was captured through 2 distinct features: a log-transformed continuous variable to measure overall length and a binary indicator for posts exceeding 140 characters. This design accounts for Weibo's historical character limits and the platform's current interface mechanism, in which longer posts are truncated and require users to click "expand" to view the full content—an additional interaction step that may introduce friction and influence user engagement behaviors.

To capture emotional attributes, sentiment polarity for the full corpus ($n=93,799$) was first classified into 3 categories (positive, neutral, and negative) using the Gemma4:e4b LLM (Google DeepMind) through prompt-based annotation. To evaluate the reliability of this automated annotation, we then constructed a silver-standard reference dataset—a high-quality reference set generated primarily through automated multimodel consensus, with human adjudication reserved for cases of full disagreement—through stratified sampling: 500 posts were randomly drawn from each of the 3 Gemma4:e4b-classified categories (1500 in total) to ensure adequate evaluation coverage across all sentiment classes, including the typically less frequent ones. These 1500 posts were independently reannotated using identical prompts by 3 Chinese LLMs—Doubao (Doubao-Seed-2.0-pro-260215 and ByteDance), Kimi (Kimi-K2.5-Thinking and Moonshot

AI), and Qwen (Qwen3.5-397B-A17B-Thinking and Alibaba Cloud)—which had no access to the Gemma4:e4b labels. Intermodel consistency among the 3 evaluator models was assessed using the ordinal Krippendorff α coefficient, and final silver-standard labels were determined through majority voting, with human adjudication applied in cases of complete disagreement among the 3 models. Agreement between the Gemma4:e4b annotations and the silver-standard labels was subsequently evaluated using the Quadratically Weighted Cohen κ (QWK) coefficient, which accounts for the ordinal structure of sentiment polarity.

Finally, to capture temporal and contextual characteristics of posting behavior, we engineered categorical features representing posting hour, season, weekend status, and posting device source, which on Weibo is publicly visible and reflects the operational context of the posting behavior (eg, mobile devices vs professional web management clients). To reduce feature sparsity and avoid overfitting to infrequent platforms, only the 10 most common device sources were retained as independent categories, while all remaining long-tail sources were grouped into an "Other" category.

Regarding the target variables (reposts, comments, and likes), social media engagement metrics typically exhibit highly skewed long-tail distributions. To stabilize variance and reduce the disproportionate influence of extreme outliers, all engagement counts were transformed using a base-10 logarithmic transformation ($\log_{10}(x+1)$). These transformed engagement metrics served as the continuous target variables for the regression models.

Model Optimization and Validation

We selected XGBoost as the primary modeling approach because gradient-boosted decision trees are well suited to capturing complex interactions and diminishing returns common in diffusion processes [58]. In addition, prior benchmarks on tabular prediction tasks have shown that tree-based ensembles often provide strong accuracy and computational efficiency at this scale [59]. As reposts, comments, and likes represent different modes of user engagement, model optimization was conducted independently for each target variable. Using the Optuna framework, we performed independent hyperparameter optimization (eg, tuning learning rate, max depth, and subsample) to minimize the validation root-mean-squared error (RMSE) for each target variable, yielding 3 distinct models. To prevent data leakage from users with multiple posts, these models were evaluated using a rigorous 5-Fold GroupKFold Cross-Validation strategy, grouped by user ID. Model performance was quantified using R^2 and RMSE.

Interpreting Feature Contributions Using SHAP

To interpret the fitted regression models, we used SHAP, a game-theoretic framework that decomposes continuous predictions into feature-level contributions [56]. SHAP supports 2 complementary forms of explanation in our study: (1) global feature importance, which identifies the most influential predictors across the corpus; and (2) dependence

plots, which visualize nonlinear effects and reveal potential thresholds where a predictor's marginal contribution changes sharply [57]. Together, these analyses provide an interpretable account of what is associated with engagement magnitude and diffusion in health-related rumor debunking posts.

Ethical Considerations

All data were collected exclusively from publicly accessible posts on the Sina Weibo platform. The study involved retrospective analysis of publicly available content and did not include any direct interaction with or intervention involving human participants. To protect user privacy, analyses were conducted at an aggregate level, and any personally identifiable information was removed or anonymized prior to analysis.

Results

Two-Stage Classification Results

We constructed the final analytical sample using a 2-stage classification pipeline. In stage 1 (debunking identification), SetFit with the Chinese-roberta-wwm-ext backbone achieved the best performance among 11 candidate models, with a weighted F_1 -score of 0.8973, indicating strong discriminative performance (Table 1). Applying this model to the 413,540 candidate posts identified 377,520 rumor debunking posts.

Table 1. Performance comparison of classification models for debunking identification (stage 1).

Model	Accuracy	Precision	Recall	F_1 -score (macro)	F_1 -score (weighted)
Traditional machine learning					
SVM ^a	0.7000	0.8971	0.7262	0.5888	0.7342
Random forest	0.6100	0.8571	0.6429	0.4994	0.6594
XGBoost ^b	0.6100	0.8814	0.6190	0.5215	0.6614
Logistic regression	0.8100	0.8421	0.9524	0.4945	0.7661
Standard deep learning					
Att-BLSTM ^c	0.6100	0.9245	0.5833	0.5481	0.6618
bert ^d -base-chinese	0.8800	0.9186	0.9405	0.7647	0.8767
Chinese-roberta ^e -wwm-ext	0.9000	0.9302	0.9524	0.8039	0.8973
Few-shot learning					
SetFit ^f (Chinese-bert-wwm-ext)	0.8900	0.9195	0.9524	0.7782	0.8853
SetFit (bert-base-chinese)	0.8900	0.9101	0.9643	0.7645	0.8814
SetFit (paraphrase-multilingual-MiniLM-L12-v2)	0.8900	0.9195	0.9524	0.7782	0.8853
SetFit(Chinese-roberta-wwm-ext) ^g	0.9000	0.9302	0.9524	0.8039	0.8973

^aSVM: support vector machine.

^bXGBoost: Extreme Gradient Boosting.

^cAtt-BLSTM: Attention-based Bidirectional Long Short-Term Memory.

^dBERT: Bidirectional Encoder Representations from Transformers.

^eRoBERTa: a robustly optimized BERT pretraining approach.

^fSetFit: Sentence Transformer Fine-Tuning.

^gBest performing model overall.

In stage 2 (health topic classification), the same SetFit (Chinese-roberta-wwm-ext) configuration again performed best, reaching a weighted F_1 -score of 0.9405, reflecting

robust classification performance (Table 2). When applied to the 377,520 debunking posts, it yielded 93,799 health-related posts (24.8%) and 283,721 non-health-related posts (75.2%).

Table 2. Performance comparison of classification models for health topic classification (stage 2).

Model	Accuracy	Precision	Recall	F_1 -score (macro)	F_1 -score (weighted)
Traditional machine learning					
SVM ^a	0.7500	0.6190	0.4333	0.6710	0.7355
Random forest	0.6600	0.4412	0.5000	0.6094	0.6656
XGBoost ^b	0.6100	0.3953	0.5667	0.5793	0.6248
Logistic regression	0.7600	0.6500	0.4333	0.6800	0.7440
Standard deep learning					
Att-BLSTM ^c	0.7000	0.5000	0.2667	0.5765	0.6680
bert ^d -base-chinese	0.9200	0.8667	0.8667	0.9048	0.9200
Chinese-roberta ^e -wwm-ext	0.9100	0.8387	0.8667	0.8939	0.9104

Model	Accuracy	Precision	Recall	F_1 -score (macro)	F_1 -score (weighted)
Few-shot learning					
SetFit ^f (Chinese-bert-wwm-ext)	0.9200	0.8929	0.8333	0.9029	0.9192
SetFit (bert-base-chinese)	0.9100	0.8889	0.8000	0.8896	0.9086
SetFit (paraphrase-multilingual-MiniLM-L12-v2)	0.9400	0.9615	0.8333	0.9256	0.9387
SetFit (Chinese-roberta-wwm-ext) ^g	0.9400	0.8750	0.9333	0.9299	0.9405

^aSVM: support vector machine.
^bXGBoost: Extreme Gradient Boosting.
^cAtt-BLSTM: Attention-based Bidirectional Long Short-Term Memory.
^dBERT: Bidirectional Encoder Representations from Transformers.
^eRoBERTa: a robustly optimized BERT pretraining approach.
^fSetFit: Sentence Transformer Fine-Tuning.
^gBest performing model overall.

Temporal and Source Characteristics of Health-Related Rumor Debunking on Weibo

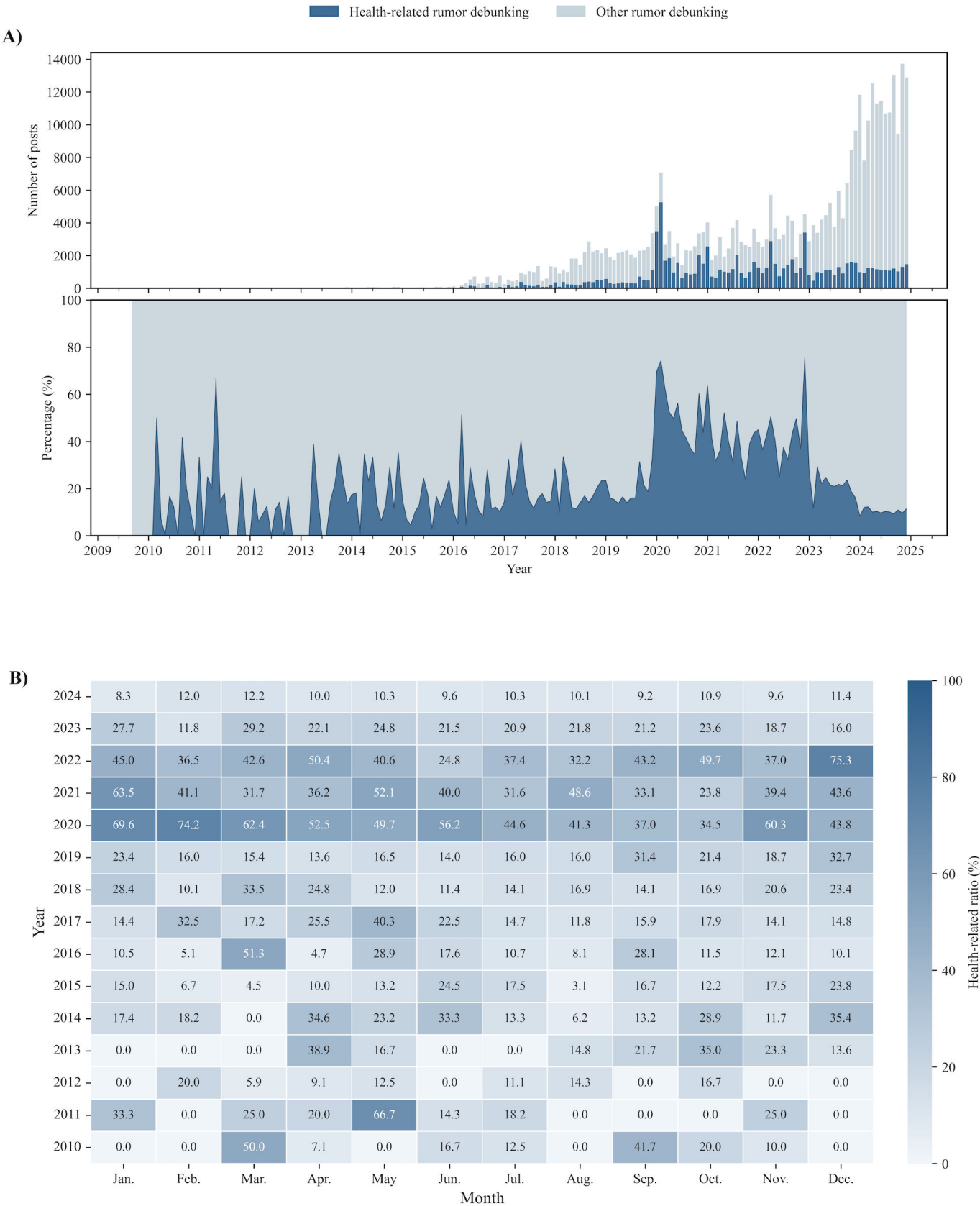
Using the health-related debunking posts identified by the 2-stage pipeline, we examined temporal trends and source composition from 2009 to 2024.

Temporal Patterns of Health-Related Rumor Debunking

From 2009 to 2024, the overall volume of rumor debunking posts increased substantially, with 2 pronounced peaks

in 2020 and 2024 (Figure 2A). Importantly, these peaks differed in topical composition. The 2020 peak was characterized by an unprecedented concentration of health-related debunking posts, temporally aligned with the outbreak of the COVID-19 pandemic. In contrast, the 2024 peak was largely composed of non-health-related topics, indicating that the overall expansion of debunking activity did not translate into a parallel increase in health-related content.

Figure 2. Temporal patterns of rumor debunking posts on Sina Weibo. (A) Annual post volume and the percentage of health-related content (2009-2024). (B) Heatmap of the monthly intensity of health-related content (2010-2024). Data from 2009 were excluded from the heatmap due to incomplete temporal coverage (available only from August to December).



To further clarify how health-related debunking responded to major public health events, we examined monthly variation in the share of health-related posts (Figure 2B). The proportion rose sharply to 69.6% in January 2020 and 74.2% in February 2020, corresponding to the initial outbreak period. A second major spike occurred in December 2022 (75.3%), coinciding

with the nationwide relaxation of COVID-19 containment policies and heightened public concern about infections and medical supplies. Although the largest spikes clustered in winter months, elevated proportions were also observed in other periods (eg, April 2022: 50.4%; May 2021: 52.1%; and May 2011: 66.7%), indicating that intensified health-related debunking was not confined to winter months and is unlikely to be explained by fixed seasonality alone.

Beyond these event-driven surges, we observed a structural shift in baseline activity following the COVID-19 outbreak. As shown in Figure 2A, health-related debunking remained consistently higher during 2020 to 2024 than in the prepandemic period (2009 to 2019), indicating sustained activity in this domain throughout the postpandemic period.

Despite this sustained volume, the relative predominance of health topics did not persist into 2024. Although total rumor debunking activity reached its highest level in that year, the health-related share declined markedly and remained between 8.3% and 12.2% throughout the year. While the

absolute volume remained significant, in relative terms, the proportion approached the lower levels observed during the prepandemic period (2014-2019). Taken together, these results suggest that while rumor refutation remained highly active overall, the platform’s debunking emphasis shifted away from pandemic-centered health topics in 2024.

Sources of Rumor Debunking Posts

To characterize the actors involved in health-related rumor debunking, we first examined the verification types of publishers within the health domain (Table 3). Government-verified accounts accounted for the largest share of health-related debunking, contributing 43,207 (64.69%) posts. Media accounts ranked second (n=10,126, 15.16%), followed by institutions (n=5632, 8.43%) and enterprises (n=5380, 8.05%). Campus and nonprofit organizations accounted for only a marginal share, together contributing less than 4% of health-related posts.

Table 3. Comparison of verified user types between health-related and non-health-related rumor debunking posts^a. Posts without valid verification-type labels (n=95,481) were excluded from this analysis.

Verified user type	Health-related posts (n=66,796), n (%)	Non-health-related posts (n=215,243), n (%)	Total (N=282,039), n (%)
Government	43,207 (64.69)	127,824 (59.39)	171,031 (60.64)
Media	10,126 (15.16)	51,391 (23.88)	61,517 (21.81)
Institution	5632 (8.43)	20,039 (9.31)	25,671 (9.10)
Enterprise	5380 (8.05)	14,841 (6.89)	20,221 (7.17)
Campus	2438 (3.65)	1119 (0.52)	3557 (1.26)
Nonprofit	13 (0.02)	29 (0.01)	42 (0.01)

^aThe distribution of user types differs significantly between health-related and non-health-related topics ($\chi^2_1=599.5$, $N=282,039$; $P<.001$; Cramér $V=0.05$).

To contextualize this source structure, we compared it with the distribution observed in non-health-related rumor debunking. Government accounts also constituted the largest contributor outside the health domain (59.39%). Although the overall source hierarchy was similar across domains, government output was more concentrated in health-related debunking: the government-to-media ratio was approximately 4.3 in health-related posts versus 2.5 in non-health-related posts.

We further examined whether the share of government-verified publishers differed between health-related and non-health-related rumor debunking posts by conducting a Pearson chi-square test on publisher type (government vs nongovernment) across the 2 domains. The difference was statistically significant ($\chi^2_1=599.5$, $n=282,039$; $P<.001$); however, the effect size was small (Cramér $V=0.05$), suggesting that government accounts were only modestly more prevalent in health-related debunking than in non-health-related debunking.

Topic Modeling and Temporal Patterns

Topic Modeling Results

Topic modeling was conducted on 93,797 health-related rumor debunking posts after text preprocessing. BERTopic

identified 352 topics. Sensitivity analysis confirmed that the exclusion of outlier documents did not systematically distort temporal patterns: the monthly post counts of outliers and retained posts were highly synchronized (Spearman $\rho=0.95$; $P<.001$), and their monthly distributions were statistically indistinguishable (Kolmogorov-Smirnov $D=0.04$; $P=.99$). Quantitative evaluation further demonstrated high clustering quality on the retained documents, yielding a topic coherence score of 0.6589 and a topic diversity score of 0.8043. As established measures of semantic consistency [60] and topic distinctiveness [61], these scores exceed the common empirical reference values of 0.55 and 0.70, respectively, indicating strong semantic consistency and high topic distinctiveness. Building upon these mathematically robust clusters, our human-in-the-loop framework generated the final interpretable labels. Table 4 reports the 10 most prevalent topics, including post counts, corpus proportions, and representative keywords (translated).

Table 4. The 10 most prevalent topics identified from health-related rumor debunking posts on Sina Weibo^a.

Rank	Topic	Posts, n (%)	Top keywords (translated)
1	Vaccine	1140 (1.22)	Vaccination, vaccine, first, dose, Guangzhou, injection, second, later, no longer, month
2	Debunking report	1031 (1.10)	Zhao (surname), police, according to law, impersonation, administrative detention, epidemic-related, investigation, spread, egregious, reselling
3	Personal care	867 (0.92)	Skin, hair loss, cosmetic facial mask, hair follicles, hair, sunscreen, skin care, silicone oil, shampoo, hair washing
4	Sleep hygiene	733 (0.78)	Sleep, staying up late, nap, falling asleep, exercise, before bed, sleeping well, regular, sleep enough, quality
5	Food myths	721 (0.77)	Food incompatibility, food, 1935, widespread, everything is normal, discard, alarmist, peanut, animal, chestnut
6	PPE ^b myths	704 (0.75)	Protective mask, pulmonary nodules, ethylene oxide, causing, Ye Xianwei, doubts, Guizhou Province, clarification, long-term, medical department
7	Local misinformation	659 (0.70)	Longnan, Gansu, municipal party committee, cyberspace administration, all-out, online rumors, makeshift hospital, Wuhan, Chengdu Middle School, Shijiazhuang
8	Communicable diseases	656 (0.70)	Novel, coronavirus, all, pneumonia, layered, prevent, antiviral, vinegar fumigation, prevention, salt water
9	Nutrition and lifestyle myths	653 (0.70)	Weight loss, skipping meals, dinner, eating dinner, obesity, carbohydrates, diet, autophagy, staple food, daily
10	Vision and eye health myths	619 (0.66)	Myopia, eyes, National Eye Care Day, prescription, vision, surgery, wearing glasses, presbyopia, glasses, myopia

^aTopics are ranked in descending order of post count. Representative keywords were translated from Chinese. Under the current BERTopic settings, 46,616 (49.70%) posts were classified as topic –1 by Hierarchical Density-Based Spatial Clustering of Applications with Noise (outliers) and were excluded from topic-specific clusters.

^bPPE: personal protective equipment.

Overall, topic prevalence was highly dispersed rather than dominated by a small set of themes. Even the most frequent topic (“vaccine”) accounted for only 1.22% of the corpus, and the remaining top-ranked topics each contributed similarly small shares. To facilitate interpretation, we grouped the top 10 topics into 2 broader domains: public health and crisis-related governance, as well as lifestyle and routine health management.

The first domain centers on infectious diseases, preventive measures, and crisis communication. For example, “vaccine” (rank 1) includes terms related to vaccination schedules and dosage, while “communicable diseases” (rank 8) and “personal protective equipment (PPE) myths” (rank 6) capture discussions involving pathogens and protective practices. In addition, some clusters, such as “debunking report” (rank 2) and “local misinformation” (rank 7), feature governance- and enforcement-related keywords (eg, “police,” “investigation,” and place names), suggesting that part of health-related rumor debunking is embedded within broader information governance and administrative enforcement contexts.

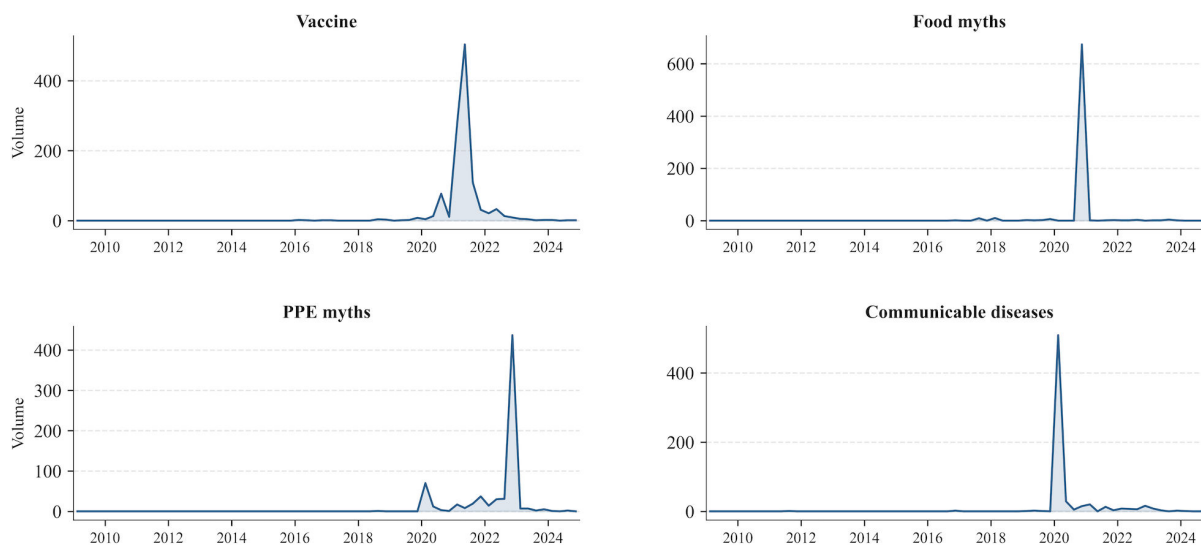
The second domain reflects recurring concerns in everyday health management. “Personal care” (rank 3) and “sleep hygiene” (rank 4) relate to appearance- and lifestyle-oriented issues, while diet-related misinformation appears in both “food myths” (rank 5) and “nutrition and lifestyle myths” (rank 9). “Vision and eye health myths” (rank 10) focuses on myopia correction and eye protection. Taken together, these results indicate that health-related rumor debunking on Weibo spans both crisis-oriented public health topics and a wide range of routine concerns about personal well-being.

Temporal Patterns of Major Topics

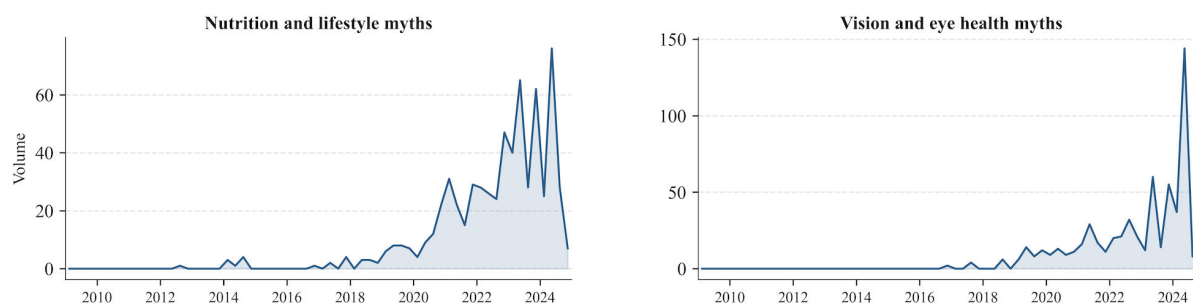
Figure 3 presents the temporal trajectories of post volume across the top 10 topics. Topic activity remained relatively low before 2019 but increased markedly thereafter. Overall, the temporal dynamics of these topics fall into 3 archetypes that reflect different attention-mobilization patterns in health-related debunking: event-driven spikes, sustained growth, and recurrent fluctuations.

Figure 3. Temporal patterns of the top 10 health-related rumor debunking topics on Weibo. The time axis begins in 2010 because the earliest post among the top 10 topics was published on March 8, 2010. PPE: personal protective equipment.

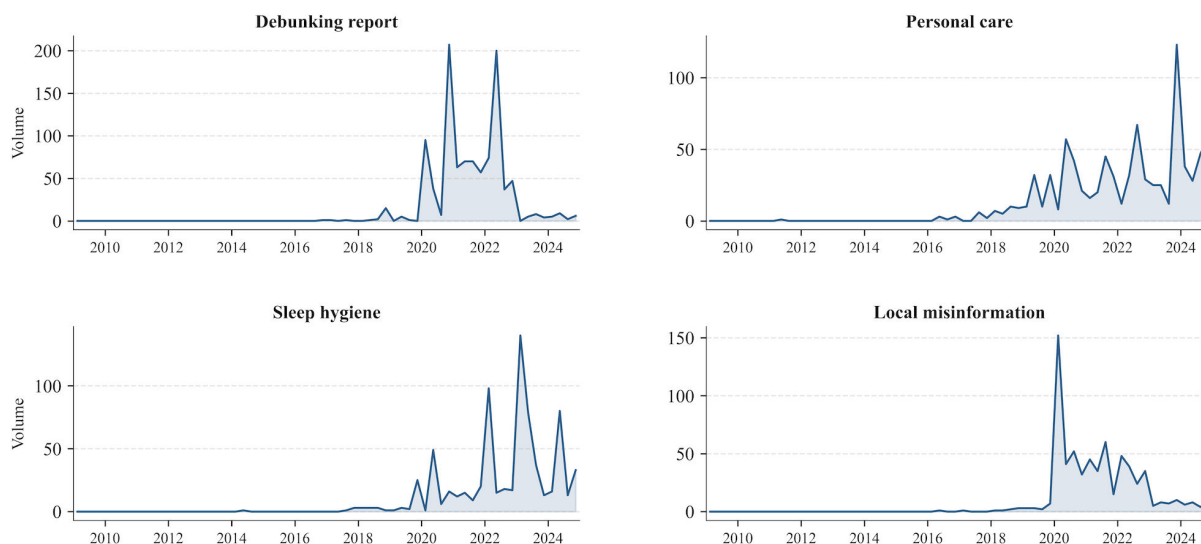
A) Event-driven types



B) Sustained growth types



C) Recurrent fluctuation types



The first archetype, event-driven spikes (eg, vaccine, food myths, PPE myths, and communicable diseases), is defined by long periods of low activity punctuated by short-lived, high-intensity surges. These topics respond sharply to external shocks and then quickly regress toward baseline levels. For instance, vaccine-related debunking remained marginal before 2020 but rose steeply during the pandemic period, peaking around mid-2021 before declining again. Similarly, PPE myths and communicable diseases exhibit synchronized spikes in early 2020, consistent with the early phase of COVID-19–related uncertainty. Food myths differ slightly in that it shows a concentrated, single-pulse peak in 2021 rather than repeated waves.

In contrast, the second archetype, sustained growth (eg, nutrition and lifestyle myths and vision and eye health myths), shows a gradual upward trajectory with mild fluctuations, rather than abrupt shock-driven surges. These topics increase gradually over time, suggesting an expanding and persistent demand for correction in routine health management. Vision and eye health myths is particularly illustrative, rising from near-zero levels around 2016 and continuing upward to its highest activity in 2024, indicating steadily growing attention in debunking practice.

The third archetype, recurrent fluctuations (eg, debunking report, personal care, sleep hygiene, and local misinformation), differs from the other 2 patterns in that attention is repeatedly reactivated rather than driven by a single shock or a gradual rise. These topics show irregular volatility with multiple recurring peaks, reflecting episodic bursts of debunking activity over time. Although their overall activity increased after 2019, the repeated rises and declines suggest that attention cycles are triggered intermittently rather than accumulating steadily.

Analysis of Extracted Information

Overall Frequencies of Mentioned Groups

Table 5 presents the distribution of target audiences identified through post-level targeting feature analysis. The majority of posts were classified as either nonoriented or universally oriented, accounting for 91.31% (n=85,648) of the 93,799 posts for gender and 87.64% (n=82,205) for age. This indicates that most health rumor debunking content was not explicitly tailored to specific demographic groups.

Table 5. Distribution of specific target audiences in health rumors^a.

Demographic characteristics	Count, n (%)
Gender	
Female	5352 (5.71)
Male	2799 (2.98)
Age group	
Minor (<18 y)	4878 (5.20)
Youth (18-44 y)	5020 (5.35)
Middle aged adults (45-59 y)	4058 (4.33)
Older adults (≥60 y)	3590 (3.83)

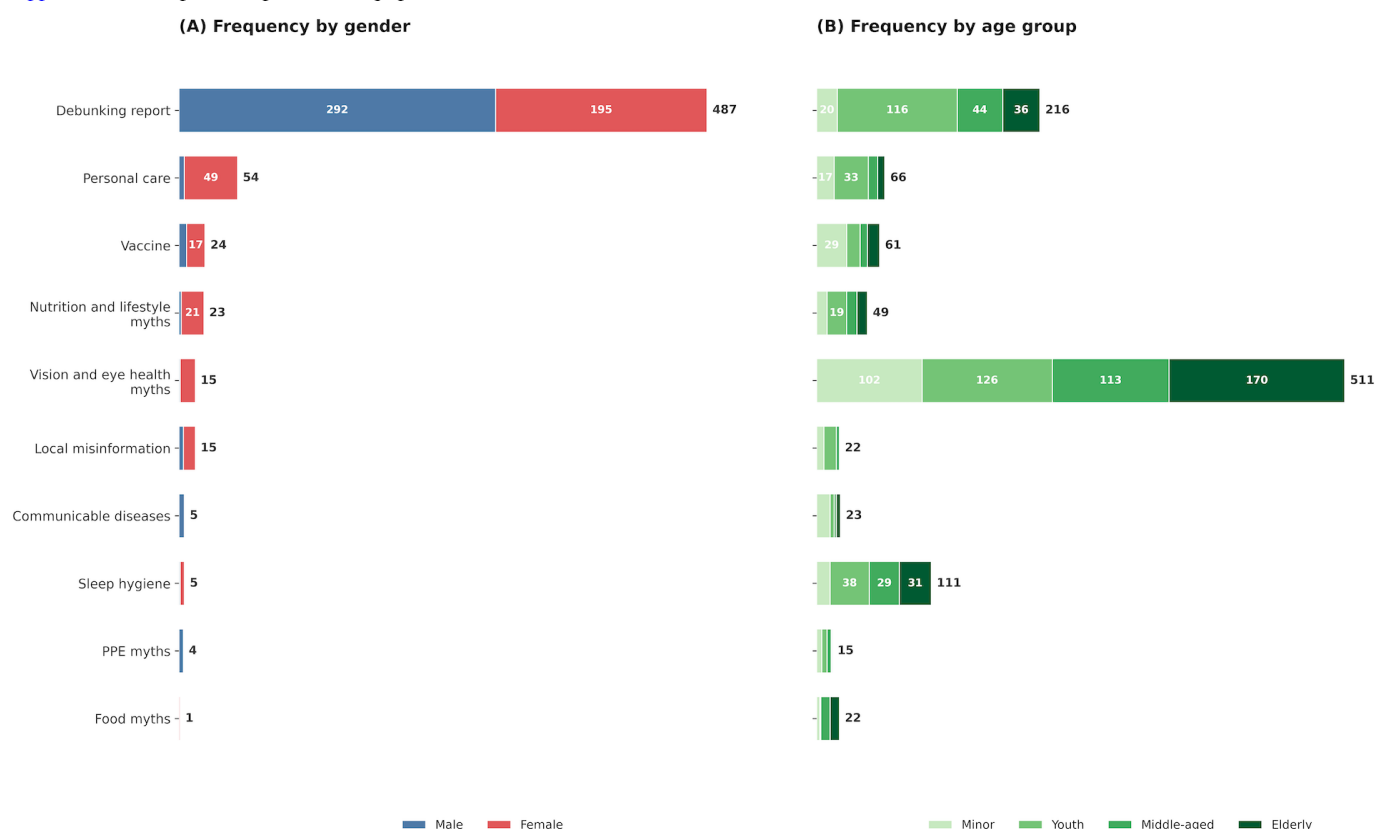
^aAge groups were categorized based on legal definitions and epidemiological evidence. Statistics are based on unique posts. For each demographic category, the count represents the number of unique posts containing a specific targeting feature. A single post may mention or target multiple age groups and therefore may be counted in more than one specific age-group category. Posts with no specific demographic features (“nonoriented”) or with all demographic groups mentioned (“universally oriented”) were grouped together and excluded from the specific demographic categories.

Among posts exhibiting explicit demographic orientation, clear disparities were observed. Female-oriented posts (n=5352, 5.71%) substantially outnumbered male-oriented posts (n=2799, 2.98%). With respect to age, youth constituted the most frequently targeted group (n=5020, 5.35%), followed by minors (n=4878, 5.20%) and middle-aged adults (n=4058, 4.33%). In contrast, older adult-oriented posts were the least prevalent among the defined age categories (n=3590, 3.83%).

Associations Between Health Content and Specific Groups

Figure 4 illustrates the distribution of gender- and age-referenced mentions across the top 10 BERTopic-derived themes in health-related debunking posts. Overall, both demographic dimensions exhibited marked topic-level variation, indicating that references to specific populations were unevenly distributed across thematic contexts.

Figure 4. Frequency of group mentions in different topics. The total frequency of mentions per topic is shown to the right of each bar, while segment labels indicate specific subgroup counts. Labels are omitted for smaller segments for visual clarity. Detailed statistics are provided in [Multimedia Appendix 2](#). PPE: personal protective equipment.



In terms of gender, the aggregate frequency of mentions across the top 10 topics was relatively balanced between men ($n=321$) and women ($n=312$). However, distinct disparities emerged at the thematic level. The topic “debunking report” accounted for the largest share of gendered references for both genders, with a higher frequency in posts mentioning men ($n=292$) than women ($n=195$). In contrast, lifestyle-related topics showed a pronounced skew toward female mentions. “Personal care” was overwhelmingly associated with women ($n=49$) compared with men ($n=5$), and “nutrition and lifestyle myths” similarly exhibited a female-dominant pattern (women: $n=21$ and men: $n=2$). Conversely, men appeared slightly more frequently in technical or acute health topics such as “communicable diseases” and “PPE myths,” although the absolute numbers for these categories were low.

Age-related references displayed a more differentiated thematic distribution. Overall, youth ($n=367$) and older adults ($n=280$) were the most frequently mentioned age groups. “Vision and eye health myths” emerged as the most significant topic in the context of age-specific mentions, accumulating the highest aggregate frequency ($n=511$). This topic showed broad relevance across the lifespan, ranging from the older adults ($n=170$) and youth ($n=126$) to the middle-aged adults ($n=113$) and minors ($n=102$), indicating that eye health is a concern shared across age groups rather than confined to any single generation. The “youth” group dominated the “debunking report” category ($n=116$), significantly outpacing other age groups. In addition, “personal care” was primarily associated with youth ($n=33$) and minors ($n=17$), with

comparatively few references to the older adults ($n=7$). The older adult group showed notable representation in “sleep hygiene” ($n=31$), ranking second only to youth ($n=38$).

Comparing gender- and age-based references reveals differences in how demographic attributes were distributed across topics. “Vision and eye health myths” accounted for a substantial proportion of age-related mentions but contained relatively few gender-specific references (total gender mentions=15). In contrast, “debunking report” posts exhibited a high density of gender references ($n=487$) while containing comparatively fewer age-specific mentions than the vision-related topic. Together, these patterns indicate that age and gender were emphasized in different thematic contexts within health-related debunking posts.

The Association Between Groups and Specific Topics

To characterize health narratives associated with different populations, we analyzed the most salient keywords linked to each demographic group ([Multimedia Appendix 3](#)). Across all strata, pandemic-related terms—most notably “pandemic,” “pneumonia,” and “infection”—were consistently prominent, establishing a shared thematic baseline for health rumor debunking during the study period. Beyond this common context, distinct gender- and age-specific keyword patterns were observed.

The female-associated keyword profile was dominated by terms related to reproductive and maternal health,

including “uterus,” “ovary,” “gestational period,” “fetus,” and “pregnancy brain.” In contrast, the male-associated profile exhibited a markedly different focus, featuring keywords related to acute cardiovascular events and pandemic control measures, such as “cardiac arrest,” “heart,” “choking,” “lockdown,” and “quarantine.” Reproductive health-related terms were largely absent from the male profile.

For minors, the dominant keywords clustered around conditions and care-related contexts, with terms such as “myopia” and “leukemia” co-occurring with caregiving-related keywords, including “injection,” “taking medicine,” and “milk.”

The youth group displayed a heterogeneous keyword set combining lifestyle-related behaviors—such as “cola,” “eating hotpot,” and “staying up late”—with terms denoting severe acute health outcomes, including “sudden death,” “myocardial infarction,” and “esophageal cancer.”

For the middle-aged group, prominent keywords clustered around obstetric and reproductive events (eg, “premature birth,” “hemorrhage,” and “gave birth”), alongside general health maintenance terms such as “soybean” and “exercise.”

In contrast, the older adult keyword profile was dominated by pandemic-related and prevention-oriented terms, including “COVID-19,” “testing,” “vaccine,” “vaccination,” “pneumonia,” “virus,” and “infection.” Alongside these, keywords reflecting age-specific vulnerability and support contexts—such as “heatstroke,” “decline,” “function,” and “medical insurance”—were also prominent.

Distribution of Social Roles and Identity Labels

While gender and age provide a structural overview of demographic targeting, the distribution of social roles reveals more context-specific forms of situational vulnerability (see [Multimedia Appendix 4](#) for the top 20 most frequent social roles). Social role mentions were distributed across several distinct categories. Generic identity labels were most prevalent, with “netizens” (n=3484) and “citizens” (n=3432) appearing most frequently.

Among specific social identities, “pregnant women” (n=1482) and “students” (n=1473) emerged as the most prominent groups, followed by “patients” (n=1298) and “parents” (n=1220).

Professional roles were also prominently represented, with “experts” (n=2659) ranking third overall and “physicians” (n=1479) appearing at high frequency. Together,

these patterns suggest that, beyond generalized references to the public, health-related narratives are frequently anchored in professional authority as well as roles associated with reproduction, education, and caregiving.

Determinants of Diffusion in Health-Related Rumor Debunking Posts

Validation of Sentiment Feature Extraction

To examine the potential impact of sentiment on information diffusion, we first validated the reliability of the extracted sentiment features. The internal consistency among the 3 evaluator LLMs used to construct the silver-standard dataset was substantial (ordinal $\alpha=0.694$; see [Multimedia Appendix 5](#) for comprehensive agreement metrics, including Fleiss κ and pairwise QWK). This confirms a sufficient and reliable level of consensus in evaluating the complex sentiment polarities of health-debunking discourse. Evaluated against this silver standard, the Gemma-4 annotations achieved a macro- F_1 -score of 0.629, an overall accuracy of 0.639, and a QWK of 0.576. Despite the inherent subjectivity and extreme semantic complexity of Weibo texts, these metrics collectively indicate a reliable and robust level of accuracy for the automated sentiment classification.

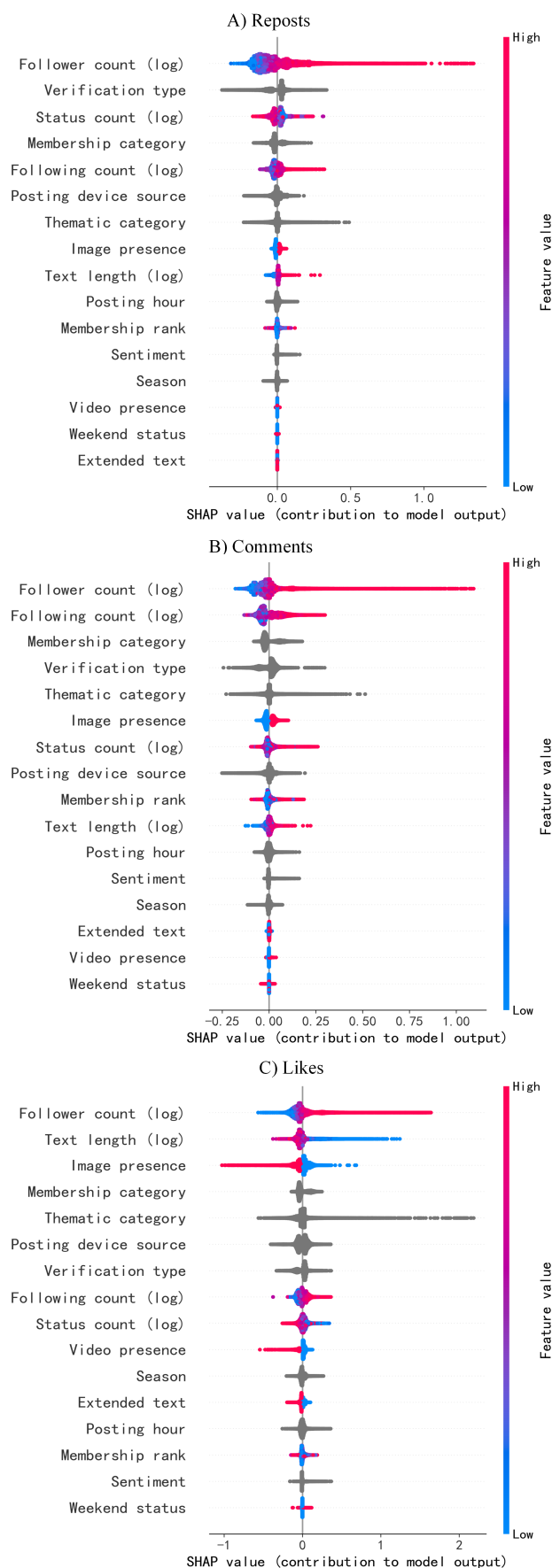
Predictive Performance of the XGBoost Regressors

The XGBoost regression models showed moderate explanatory power across all 3 engagement dimensions, consistent with the high inherent variability of social media engagement data. Specifically, the reposts model achieved an R^2 of 0.225 (RMSE=0.344), the comments model yielded an R^2 of 0.210 (RMSE=0.335), and the likes model obtained an R^2 of 0.263 (RMSE=0.572). These results indicate that the models captured meaningful structural and contextual signals associated with user interactions.

Global Feature Importance and Divergent Engagement Associations

[Figure 5](#) presents the SHAP summary plots for reposts, comments, and likes. Across all 3 dimensions, follower count (log-transformed) consistently ranked as the most important feature. The SHAP values indicate a direct positive association: accounts with larger follower counts were associated with higher engagement predictions, whereas accounts with smaller follower counts were associated with lower predictions.

Figure 5. Shapley Additive Explanations (SHAP) summary plots for feature contributions to engagement predictions: (A) reposts, (B) comments, and (C) likes. Features are ranked by mean absolute SHAP value; each dot represents one post (red=high feature value, blue=low). “Extended text” is a binary indicator of whether a post exceeds 140 characters (and is thus truncated with a “show more” link); “weekend status” indicates whether the post was published on a weekend.



Regarding content attributes, text length and image inclusion showed divergent associations depending on the mode of interaction. In the reposts model (Figure 5A) and comments model (Figure 5B), these features showed a positive directional association: longer texts and the presence of attached images were associated with positive marginal contributions to the predicted engagement. Conversely, this dynamic inverted in the likes model (Figure 5C): shorter texts and the absence of images were associated with higher predicted like counts. Notably, in this model, text length and image presence also emerged as the second and third most influential features, respectively, indicating that content-level attributes played a more central predictive role for likes than for the other engagement types.

Regarding other source characteristics, beyond follower count discussed earlier, the most prominent account metadata predictors differed across the 3 engagement models. In the reposts model (Figure 5A), identity verification status emerged as the second most important feature overall; its SHAP values exhibited wide dispersion on both sides of the zero axis, indicating that different verification categories were associated with either positive or negative marginal contributions to the prediction. In the same model, status count showed a predominantly negative directional association, with higher status counts corresponding to lower

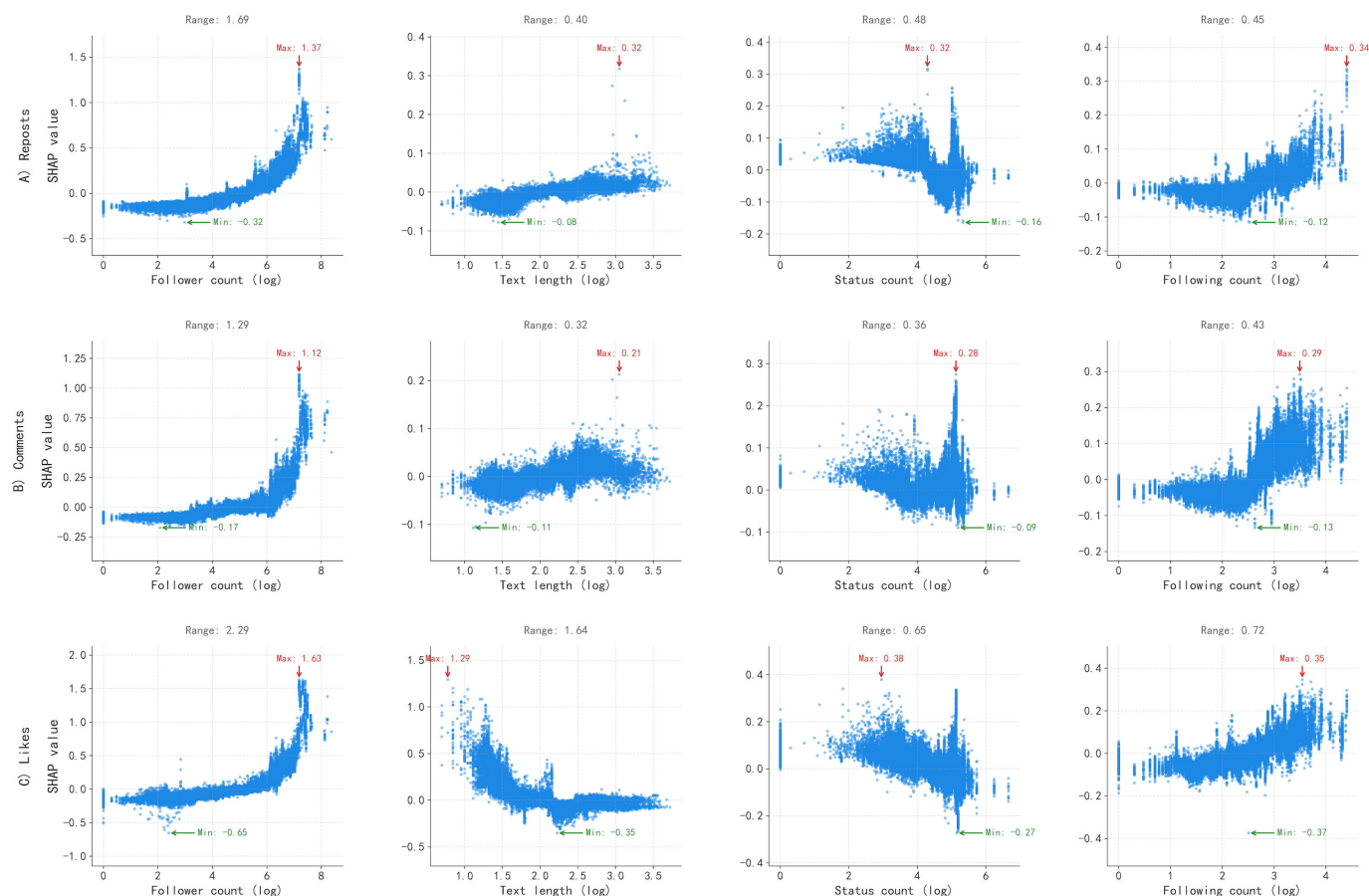
predicted repost magnitude. In the comments model (Figure 5B), following count emerged as the second most prominent feature overall, showing a positive association with predicted comments. In the likes model (Figure 5C), in contrast, no single additional source feature stood out as dominant, with the remaining source characteristics collectively showing moderate predictive importance.

Across all 3 models, temporal variables (season, time of day, weekend status) and text sentiment polarity consistently ranked in the lower tier of the feature hierarchy. Their SHAP values clustered tightly around the zero axis with minimal horizontal dispersion, indicating that sentiment polarity and posting time were associated with limited marginal contributions to final engagement outcomes.

Nonlinear Marginal Patterns of Key Predictors

Guided by their prominent roles associated with specific engagement types as identified in the summary plots, 4 continuous features—follower count, text length, status count, and following count—were selected for detailed nonlinear analysis. As illustrated in Figure 6, these dependence plots show how the marginal contribution of each feature shifts dynamically across its value distribution.

Figure 6. Shapley Additive Explanations (SHAP) dependence plots showing the nonlinear patterns of 4 key continuous features in XGBoost (Extreme Gradient Boosting) engagement predictions: (A) reposts, (B) comments, and (C) likes. The columns from left to right represent follower count (log-transformed), text length (log-transformed), status count (log-transformed), and following count (log-transformed). For each subplot, the x-axis represents the log-transformed feature value, and the y-axis indicates the corresponding SHAP value. Custom annotations denote the maximum (red arrows) and minimum (green arrows) SHAP values, with the total SHAP value range displayed at the top of each panel. The range is calculated using exact unrounded values, so subtraction of the displayed rounded labels may result in minor discrepancies.



Across all 3 engagement dimensions, follower count exhibited a consistent nonlinear pattern characterized by a distinct threshold. For values below approximately 6.0, the SHAP values were predominantly distributed in the negative region with a relatively flat trajectory. Beyond this threshold, the marginal contributions demonstrated a sharp, positive upward trend. Among the 4 analyzed continuous features, follower count exhibited the widest predictive contribution range, particularly in the likes model (SHAP range ≈ 2.3).

The dependence plots for text length revealed contrasting nonlinear patterns across the interaction types. In the likes model (Figure 6C), the highest positive SHAP values were concentrated at the lower end of the distribution, peaking at approximately 1.3 for very short texts. As text length increased, the marginal contribution exhibited a steep downward trajectory, crossing into the negative region and reaching values around -0.4 . Conversely, the reposts and comments models (Figures 6A and 6B) displayed a different dynamic. After an initial dip in the shorter text range, the SHAP values transitioned into a positive upward trend. Notably, this positive trajectory in the comments model was not sustained; the main cluster of SHAP values began to decline after reaching a localized concentration in a mid-length range.

Status count demonstrated a nonlinear trajectory. At lower levels, the marginal contributions were predominantly positive. As the post count increased, these values declined into the negative region to form a distinct trough, before exhibiting a rebound trend at the higher end of the distribution. Notably, the onset of this downward shift occurred earlier in the likes model compared to reposts and comments.

Finally, following count exhibited a clear threshold pattern. At the lower end of the distribution, the SHAP values were predominantly negative. However, as the values surpassed a mid-range threshold (approximately 2.0-2.5 on the logarithmic scale), the scatter plots crossed the zero axis and transitioned into a positive trajectory for larger following counts.

Discussion

Principal Results

Overview

By analyzing a comprehensive dataset spanning 15 years of health-related rumor debunking posts on Sina Weibo, this study characterizes the longitudinal evolution, thematic diversity, demographic representations, and engagement

predictors of corrective communication in China's digital health ecosystem. Our findings demonstrate that health-related rumor debunking on Weibo was strongly influenced by major public health crises, particularly during the COVID-19 pandemic. At the same time, the topical landscape has progressively diversified over the past decade, with lifestyle-related and well-being-related themes showing relatively stable long-term growth. In addition, although most debunking posts targeted the general public, clear differences emerged across gender and age groups in both representation and associated health concerns, and these demographic references were unevenly linked to specific health topics. Finally, interpretable machine learning indicated that engagement was associated primarily with source-level reach, with follower count being the strongest predictor across reposts, comments, and likes, while the contribution of content attributes varied by interaction type and was most pronounced for likes; sentiment and posting time showed limited associations. Considered jointly, these findings reveal a structural asymmetry in the debunking ecosystem that single-dimension analyses would not capture: the content production side is maturing, as thematic coverage is broadening and demographic differentiation is emerging, while the distribution side remains governed by source-level reach largely irrespective of content attributes. This content distribution misalignment means that advances in what is corrected and for whom do not automatically translate into proportional communicative impact.

Crisis-Driven Evolution and Subsequent Attention Rebalancing

Our longitudinal findings (2009–2024) suggest that the COVID-19 pandemic was associated with substantial shifts in the volume of health rumor debunking on Sina Weibo, with effects that appear to have extended well beyond the acute outbreak period. The sharp escalation of health-related debunking activity in 2020 likely reflected the convergence of 2 parallel dynamics. On the one hand, the outbreak generated unprecedented levels of misinformation under conditions of high public uncertainty, concerning transmission routes, protective measures, vaccines, and treatment strategies. On the other hand, the pandemic simultaneously triggered an intensified institutional response, as government agencies and platforms rapidly expanded corrective communication efforts to stabilize the online information environment during a large-scale public health emergency.

Beyond this initial surge, the elevated level of health-related debunking activity persisted well past the most acute outbreak stages. Even after the initial outbreak phase (2020) subsided, the volume of health-related debunking remained substantially elevated throughout 2020 to 2023 compared with the prepandemic baseline. This pattern suggests that the pandemic may have functioned not only as a short-term focusing event but also as a catalyst for broader governance reconfiguration. Institutional mechanisms initially mobilized under emergency conditions—including specialized debunking accounts, cross-platform coordination, and intensified administrative intervention—appear to

have become partially embedded within routine governance practices. This pattern can be interpreted through the broader lens of Punctuated Equilibrium Theory [62], which proposes that exogenous shocks can disrupt previously stable policy equilibria and generate durable shifts in governance arrangements. However, this analysis tracks volumes of debunking posts rather than the institutional arrangements themselves; therefore, this interpretation should be regarded as suggestive rather than conclusive.

At the same time, this sustained postpandemic elevation should not be interpreted solely as evidence of continuously heightened public concern regarding health misinformation. The persistence of debunking activity may also reflect the temporary continuation of institutional mandates and governance infrastructures established during the emergency phase. In other words, the prolonged visibility of health rumor governance likely emerged through the interaction between enduring public sensitivity and sustained administrative mobilization.

By 2024, however, the relative share of health-related debunking declined markedly, despite the overall volume of rumor refutation posts reaching its historical peak. Importantly, this decline does not necessarily indicate a weakening of misinformation governance capacity. Rather, it may reflect a redistribution of institutional attention and potentially of broader public salience as pandemic-related risks gradually lost their exceptional status and competing social issues regained visibility within the broader governance agenda. Alternative explanations—including shifts in official governance mandates and potential changes in platform-level recommendation algorithms that may influence content visibility—also warrant consideration, although the present dataset does not allow us to formally separate these contributing factors. This mechanism is broadly consistent with the issue attention cycle [63], which suggests that highly salient public problems gradually lose centrality as crisis urgency fades and attention shifts toward emerging concerns. As with Punctuated Equilibrium Theory mentioned earlier, this framework is offered as an interpretive lens rather than a directly tested mechanism, given that our data capture debunking activity rather than the underlying processes of public attention themselves.

Together, these findings outline a dynamic cycle of crisis-driven rumor governance characterized by emergency escalation, partial institutionalization, and subsequent rebalancing of attention across issue domains.

State-Centered Governance in Health Rumor Debunking

Our findings reveal a predominantly state-centered configuration in China's health rumor debunking ecosystem. Government-verified accounts contributed nearly two-thirds (64.69%) of all health-related debunking posts, substantially exceeding the combined output of media, institutional, and enterprise accounts. While the share of government accounts was statistically higher in health-related debunking than in non-health-related debunking (chi-square test, $P < .001$),

the effect size was small (Cramér $V=0.05$), indicating that government dominance is a general feature of Sina Weibo's debunking ecosystem rather than a phenomenon unique to the health domain. Nonetheless, this dominance is amplified within the health context: the government-to-media output ratio reached approximately 4.3:1 in health-related debunking, compared with 2.5:1 in non-health-related debunking. This contrast suggests that, although state actors play a leading role across topic domains, their relative weight is even more pronounced when the corrective communication concerns health.

This configuration is consistent with Media System Dependency Theory [64], which posits that audience reliance on particular information sources intensifies under conditions of uncertainty and perceived risk. As health rumors often involve scientific complexity and potential collective consequences—such as panic buying, vaccine hesitancy, or heightened public anxiety—commercially oriented media, shaped by market incentives and attention dynamics, may be less inclined to function as definitive arbiters of truth. State actors, in contrast, are institutionally positioned to supply administrative authority and offer authoritative reference points that the public can draw upon when navigating conflicting health claims online.

This state-centered pattern, however, should not be read as a universal template for misinformation governance. Comparative evidence from Canada and the United States points to a more distributed, multiactor model, in which corrective communication is shared across public health agencies (eg, the Centers for Disease Control and Prevention and Health Canada), independent fact-checking organizations, professional medical associations, and academic experts, rather than concentrated within central government accounts [65,66]. Two factors may help explain this divergence. First, in contexts where institutional trust is fragmented along political or ideological lines, corrective messages delivered through a diverse set of nongovernmental and quasi-governmental actors may achieve broader credibility than messages issued directly by central authorities [67,68]. Second, in such settings, direct government involvement in online information governance can itself become politically contested, which makes distributed governance both a normative preference and a pragmatic strategy [69]. Overall, these contrasts indicate that effective debunking architectures are shaped by broader institutional and sociopolitical contexts and that the Chinese pattern documented here reflects one viable configuration rather than a generalizable blueprint.

Thematic Evolution: From Acute Survival Concerns to Everyday Well-Being

The thematic evolution of health rumor debunking suggests a gradual broadening of public attention from acute crisis-related concerns toward a wider range of everyday health and well-being issues. This shift plausibly reflects the increasing ubiquity of social media and broader changes in how health information is accessed, circulated, and interpreted online [70]. Within this maturing digital environment, we observe a clear divergence in temporal trajectories across

topic types. Event-driven topics, such as vaccines and personal protective equipment, tend to follow an “explosive but transient” pattern, with sharp spikes closely aligned with major epidemic phases. In contrast, lifestyle-oriented topics, most notably vision and eye health, as well as nutrition, exhibit more sustained growth over time, suggesting enduring relevance beyond discrete crisis periods.

This observed reorientation of public attention may be interpreted through the lens of Maslow's hierarchy of needs [71]. During the height of the pandemic, public attention was disproportionately concentrated on safety-related concerns, including infection avoidance and immediate risk reduction. As the perceived severity of the threat declined, attention appears to have gradually expanded toward issues associated with everyday functioning and quality-of-life management.

At the same time, the sustained prominence of lifestyle-related rumors can be understood in relation to the “medicalization of daily life” [72], whereby ordinary behaviors—such as screen use, diet, and sleep—are increasingly framed in medicalized terms that invite expert-like intervention. Within this context, routine health anxieties may accumulate over time, generating ongoing demand for ostensibly “scientific” optimization and creating a fertile environment for pseudoscientific claims, dietary myths, and “quick-fix” narratives. The coexistence of these distinct temporal archetypes suggests that effective debunking calls for simultaneously managing acute crisis response and sustained routine correction, a duality whose practical implications are further shaped by how corrections circulate and reach their audiences.

Demographic Differentiation in Health Narratives: Patterns Across Gender, Age, and Social Roles

Our demographic analysis suggests that digital health narratives are not neutral but systematically differentiated along gender and age lines. To interpret these disparities, we draw on several sociopsychological frameworks while remaining attentive to the boundaries of our data: because our dataset captures keyword-level textual representations rather than direct psychometric or behavioral measures, we frame these theoretical insights as exploratory interpretations rather than definitive causal mechanisms. In parallel, we situate these narrative differences within the structural realities of the digital environment—platform demographics, algorithmic biases, and contemporary digital culture.

With respect to gender, female-related mentions ($n=5352$, 5.71%) substantially outnumbered those associated with men ($n=2799$, 2.98%), and the female-associated keyword profile was dominated by terms related to reproductive and maternal health (eg, “uterus,” “ovary,” and “pregnancy”). This patterned framing could tentatively be interpreted through the lens of Objectification Theory [73], which highlights how women's bodies are often rendered salient through appearance- and function-oriented discourses. A more direct alternative explanation lies in the engagement-driven algorithmic structure of social media. Prior research

consistently demonstrates that women are more active seekers and sharers of online health information, particularly regarding maternal and routine well-being [74-76]. Drawing on this prior evidence, one possibility is that platform algorithms—which prioritize content with high interaction rates—disproportionately amplify reproductive and appearance-oriented debunking content because these topics are documented to generate engagement from female demographics. We note that this remains an inference drawn from external literature rather than from engagement metrics within our own dataset.

In contrast, male-related keywords clustered around 2 distinct categories: acute and high-severity health events (eg, “cardiac arrest” and “choking”) and pandemic control terms (eg, “lockdown” and “quarantine”). While the acute-event cluster resonates with tentative insights from Hegemonic Masculinity theory [77]—which has been used to explain the relative invisibility of routine male health concerns—it may also reflect the interaction between gendered health information behaviors and platform visibility dynamics. Prior research suggests that men generally engage less with routine or preventive health information online [78]. Drawing on prior platform studies, low-interaction topics may be less likely to receive algorithmic amplification [79], whereas acute and emotionally arousing medical emergencies are more commonly associated with rapid sharing and public attention [5]. Within this framework, routine male health concerns may remain relatively less visible while acute medical events become more prominent in rumor debunking discussions. The co-occurrence of pandemic control terms in the male-associated profile is plausibly a separate phenomenon, reflecting the broader prominence of these topics during the COVID-19 period rather than a gender-specific pattern. As mentioned earlier, these interpretations rest on external evidence rather than on engagement data within our own corpus.

Across the life course, vulnerability profiles further diverge. Youth emerged as the most frequently referenced group (n=5020, 5.35%), a pattern that closely mirrors Weibo’s user demographics, where young adults constitute the platform’s core population [80]. Beyond frequency, the youth-associated keyword profile was notably heterogeneous, including both lifestyle-related behaviors—such as “cola,” “hotpot,” and “staying up late”—and terms denoting severe acute health outcomes, such as “sudden death,” “myocardial infarction,” and “esophageal cancer.” This heterogeneous profile likely reflects the breadth of health concerns relevant to Weibo’s predominantly young user base, encompassing both everyday lifestyle topics and the acute medical incidents that periodically attract public attention in this demographic.

Minors also appeared as a highly visible group (n=4878, 5.20%), reflecting their socially recognized vulnerability and the prioritization of child protection in public discourse. The prominence of caregiving-related keywords (eg, “injection,” “taking medicine,” and “milk”) alongside the high frequency of the “parents” social role (n=1220) suggests that minors’ health concerns in this corpus are frequently mediated through caregiver discourse rather than appearing as self-expressed concerns. In line with prior research [81],

this pattern may reflect parental proxy health information seeking, in which caregivers actively engage with health-related content on behalf of children—although our keyword-level data describe textual co-occurrence rather than directly capturing information-seeking behaviors. Correspondingly, the discourse is heavily concentrated on “vision and eye health” (Figure 4) and keywords such as “myopia” and “leukemia” (Multimedia Appendix 3), reflecting both routine developmental concerns and heightened anxieties surrounding severe pediatric conditions. This pattern is further reinforced by the prominence of the “student” social role (n=1473), suggesting that health narratives about minors are frequently framed within educational settings rather than purely biological trajectories of development.

Moving further along the life course, the middle-aged group exhibits a distinctive and multifaceted keyword profile, combining high-risk obstetric terms (eg, “premature birth,” “hemorrhage,” and “gave birth”) with general health maintenance terms (eg, “soybean” and “exercise”). The presence of obstetric terms potentially indicates that reproductive risk remains a salient concern in middle-aged discourse, a pattern that may be situated within broader demographic shifts in contemporary China, including the growing prevalence of pregnancies at advanced maternal age following the relaxation of birth control policies under the 2- and 3-child initiatives. We note, however, that the present cross-sectional data do not directly link this keyword pattern to those policy shifts. The co-occurrence of health maintenance terms further suggests that middle-aged health discourse is not exclusively framed around acute reproductive risks but also encompasses general well-being concerns.

By contrast, the older adult group (n=3590) exhibited the lowest overall frequency among the defined age groups, although the gap relative to the other groups was modest. Rather than reflecting outright digital marginalization, this lower frequency may partly arise because older adults are often represented by others rather than through direct self-expression—a pattern that, although different in form, parallels the proxy-mediated representation observed for minors above, and is consistent with prior research on age-related disparities in social media use [75]. The older adult-associated keyword profile was dominated by pandemic-related terms (eg, “COVID-19,” “vaccine,” “pneumonia,” and “infection”), reflecting the prominence of older adults as a high-risk population during the COVID-19 period. Alongside these, keywords such as “heatstroke,” “decline,” “function,” and “medical insurance” suggest that older adult health narratives in this corpus also tend to be framed around environmental risks and institutional support, although our data cannot directly speak to the underlying agency dynamics.

In parallel with these demographic patterns, the high frequency of “experts” (n=2659) and “physicians” (n=1479) in the overall corpus points to the co-occurrence of professional roles alongside demographic identity labels within rumor debunking discourse. Although our keyword-level data do not directly distinguish the discursive functions these mentions serve, this co-occurrence pattern is broadly consistent with

prior observations of expert-driven health communication, in which professional voices and lay subjects of health concern jointly populate corrective discourse.

Although the majority of health rumor debunking posts in our corpus address a general audience without explicitly mentioning specific groups, the distinct patterns observed within the targeted subset suggest the potential value of precision communication. For health communication practitioners, these differentiated findings indicate that governance efforts might be optimized by supplementing standard universal broadcasts with tailored messages. When addressing niche health concerns, configuring the framing to the distinct contexts of specific genders, age groups, and social roles could enhance the relevance of corrective information. Across these patterns, we emphasize that our analysis describes systematic textual co-occurrences within rumor debunking discourse and that the underlying mechanisms—whether psychological, cultural, or algorithmic—warrant direct empirical testing in future research.

Predictors of Engagement: Source Hierarchy and Differentiated Engagement Modes

Our engagement analysis suggests a partial “source-over-content” tendency, particularly for diffusion (reposts) and interactive engagement (comments): the spread of debunking messages appears to be more strongly associated with who delivers the message than with what the message is about. Across all 3 engagement types, follower count (operationalized as the log-transformed follower count) consistently emerged as the dominant predictor, while sentiment and posting time showed comparatively limited predictive contributions. One plausible interpretation is that, in information-saturated social media environments, users often lack the time, cognitive resources, or domain-specific expertise required to systematically evaluate complex health information. Under such conditions, they may rely on readily available peripheral cues—such as audience size and verification-related signals—as heuristic indicators of credibility and relevance. This behavioral pattern is consistent with the Elaboration Likelihood Model, which describes the role of peripheral processing when motivation or ability for deep evaluation is limited [82]. We note, however, that our data permit only predictive associations and cannot directly test the underlying cognitive mechanisms; the Elaboration Likelihood Model is therefore offered as one plausible interpretive lens rather than a tested explanation.

Crucially, the associations of source and content features diverged depending on the mode of engagement. For reposts, follower count was followed by institutional verification (verified type) as the next most prominent source predictor, suggesting that diffusion breadth is more strongly associated with high-status, institutionally credentialed sources than with the substantive content of the post. For comments, in contrast, following count emerged as the second most prominent source predictor and was positively associated with predicted comments. One possible interpretation is that accounts following many others tend to be more active social participants, whose posts may correspond to environments

more conducive to dialogue. Social exchange theory [83] offers a complementary interpretive frame in which reciprocal networking signals may correspond to lower psychological thresholds for audience response [84], but this interpretation should be treated as exploratory given the observational nature of our data.

The most striking departure from this source-dominant pattern appeared in the likes model. Although follower count remained the leading predictor, text length and image inclusion rose to become the second and third most prominent features, with shorter texts and the absence of images associated with higher predicted likes. This pattern suggests that the lightweight, low-effort nature of liking may render lightweight content (brief, text-only posts) more compatible with the cognitive and behavioral cost of the engagement act itself. For likes, therefore, content-level attributes appear to play a more central predictive role than for reposts or comments, partially qualifying the broader “source-over-content” framing.

Despite these mode-specific nuances, the overarching ecosystem remains marked by a pronounced concentration pattern reminiscent of the Matthew Effect [85]. The consistent dominance of follower count across all 3 engagement metrics, together with the predominantly negative SHAP values observed for low-follower accounts, suggests a cold-start dynamic in which accounts lacking sufficient initial audience size struggle to gain visibility. Importantly, this pattern need not reflect user-level cognitive heuristics alone; it may also be shaped substantially by platform-level algorithmic mechanisms. On Weibo, recommendation and ranking algorithms tend to amplify content from accounts that have already accumulated engagement, producing a feedback loop in which visibility begets further visibility. The “source-over-content” pattern we observe in the data is therefore likely a joint product of user heuristics and platform algorithmic design, rather than a pure reflection of either mechanism alone. Without access to platform algorithmic data, we cannot fully disentangle these 2 contributing pathways, and we offer this interpretation as a hypothesis warranting further investigation.

Together, these patterns suggest a visibility disadvantage for credible but less-followed sources. As the SHAP dependence plots show, accounts with follower counts below approximately 10^6 were associated with predominantly negative marginal contributions across all 3 engagement dimensions, with a sharp positive inflection emerging only beyond this threshold. Credible but low-resource voices may therefore remain structurally below diffusion thresholds regardless of content quality, implying that effective governance requires not only producing accurate corrections but also enabling their circulation. Platform-level interventions, such as algorithmic amplification mechanisms or endorsement features that boost the visibility of credible grassroots sources, may help ensure that evidence-based correction is not structurally suppressed by deficits in social capital.

Limitations

First, contextual generalizability is constrained by the sociopolitical setting of this study. Our analysis is based on Sina Weibo within China's specific information governance framework. While this context offers a valuable lens into a relatively state-centered debunking ecology, it may limit the transferability of our findings to other institutional environments. In particular, the administrative-led configuration observed here differs from the more distributed, multiactor models discussed earlier, limiting the extent to which our findings generalize to such settings. Future comparative research across political and regulatory systems is needed to examine how governance structures shape the organization and effectiveness of rumor debunking.

Second, platform specificity may restrict the applicability of our conclusions beyond Weibo's media environment. Sina Weibo functions largely as a text-based public sphere with repost-centered diffusion dynamics. In contrast, emerging platforms such as Douyin (the Chinese counterpart of TikTok) operate with different content formats and recommendation architectures, while WeChat (Tencent) represents a more closed, private-domain communication system. Future work should adopt cross-platform designs to evaluate how media format, network structure, and algorithmic logic jointly influence the visibility and diffusion of debunking content.

Third, the corpus may be incomplete because of our reliance on keyword-based retrieval at the construction stage. Specifically, using "rumor debunking" (辟谣) as the primary query term may omit corrective messages that perform fact-checking implicitly without explicit debunking labels (eg, physicians providing corrective explanations in routine health guidance). Future research could incorporate semantic dense retrieval methods or iterative active learning pipelines to capture a broader spectrum of implicit corrective communication and improve coverage of grassroots correction practices.

Fourth, our engagement data are observational, and the SHAP attributions therefore characterize predictive contributions rather than causal effects. Although our models incorporated a comprehensive set of source, content, and contextual features, unobserved factors—such as platform-level recommendation weights or offline amplification events—may still shape diffusion dynamics. Future studies should use experimental designs or formal causal inference techniques to directly isolate the mechanisms underlying user engagement.

Conclusions

By analyzing 15 years of health-related rumor debunking posts on Sina Weibo (2009–2024), this study offers a longitudinal account of official corrective communication along 3 dimensions: its thematic evolution, the populations it references, and the factors associated with its diffusion. First, the focus of health-related debunking broadened over time, moving from acute crises—most notably COVID-19—toward a wider range of routine health and well-being concerns. Second, most debunking posts were not directed at any specific demographic group; among the minority that explicitly referenced particular populations, references were unevenly distributed across gender and age and were linked to distinct health concerns. Third, our predictive analysis indicates that engagement was associated primarily with source-level reach: follower count was the strongest predictor of reposts, comments, and likes alike, outweighing the substance of the correction. Beyond this shared pattern, the secondary predictors differed across the 3 interaction types, indicating that reposts, comments, and likes are shaped by partly distinct configurations of source and content attributes rather than by a single uniform mechanism.

Viewed jointly, these 3 dimensions reveal a structural asymmetry in the debunking ecosystem: the content side is maturing in both thematic scope and demographic differentiation, while distribution remains governed by source-level reach largely irrespective of content attributes. This asymmetry carries specific governance implications. First, the coexistence of event-driven spikes and sustained growth trajectories among health-related debunking topics indicates that crisis response capacity alone is insufficient; building long-term, routine communication capacity for lifestyle-related concerns that are steadily gaining prominence may be equally important. Second, the systematic associations between demographic groups and specific health concerns, such as reproductive health terms in female-associated content and pandemic-related terms in older adult-associated content, provide an empirical baseline that practitioners can draw upon to assess the coverage and priorities of current debunking efforts. Third, the SHAP analysis revealed a clear threshold effect for follower count, below which accounts were associated with predominantly negative marginal contributions to engagement, suggesting that credible but low-resource voices face a structural diffusion barrier; platform-level mechanisms that help such sources circulate corrections may therefore be an essential complement to relying on established authoritative accounts alone.

Acknowledgments

The authors declare that generative AI tools were used during the preparation of this manuscript under full human oversight. Specifically, Gemini and DeepSeek were used to assist with code generation and debugging for data analysis, and DeepSeek was additionally used to extract demographic mentions and health-related content from the corpus. Doubao was used to generate preliminary topic labels, and Doubao, Kimi, Qwen, and Gemma-4 (the gemma4:e4b model) were used to conduct automated sentiment classification. Gemini and ChatGPT were also used to support literature searching. In addition, Grammarly, together with Gemini, Claude, and ChatGPT, was used for grammatical and lexical refinement to enhance textual clarity. All AI-assisted outputs were checked for accuracy and appropriateness by the authors before being incorporated into the manuscript. The authors are accountable for the submitted manuscript in its entirety.

Funding

This study was supported by the Sichuan Science and Technology Program (grant 2024NSFSC1075).

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YF (lead), XT (supporting)

Data curation: YF

Formal analysis: YF (lead), CH (supporting)

Funding acquisition: XT

Investigation: YF (lead), CH (supporting), YH (supporting)

Methodology: YF (lead), CH (supporting), XT (supporting), YH (supporting)

Project Administration: XT

Resources: XT

Software: YF (lead), CH (supporting), XT (supporting).

Supervision: XT

Validation: YF (lead), CH (supporting)

Visualization: YF (lead), YH (supporting), CH (supporting)

Writing – original draft: YF (lead), CH (supporting), XT (supporting), YH (supporting)

Writing – review & editing: YF (lead), YH (supporting), XT (supporting), CH (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Validation metrics for the information extraction framework (interannotator agreement, overall model performance, and category-specific performance evaluation for gender and age variables).

[\[DOCX File \(Microsoft Word File\), 26 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Detailed statistics of gender and age group mentions across the top 10 BERTopic-derived themes.

[\[DOCX File \(Microsoft Word File\), 26 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Top 20 characteristic keywords of health-related rumor debunking posts, stratified by gender and age.

[\[DOCX File \(Microsoft Word File\), 23 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Top 20 most frequent social roles identified in health-related rumor debunking posts.

[\[DOCX File \(Microsoft Word File\), 25 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Intermodel agreement and performance metrics for sentiment classification.

[\[DOCX File \(Microsoft Word File\), 24 KB-Multimedia Appendix 5\]](#)

References

1. Kaplan AM, Haenlein M. Users of the world, unite! the challenges and opportunities of social media. *Bus Horiz*. Jan 2010;53(1):59-68. [doi: [10.1016/j.bushor.2009.09.003](https://doi.org/10.1016/j.bushor.2009.09.003)]
2. Bakshy E, Rosenn I, Marlow C, Adamic L. The role of social networks in information diffusion. Presented at: Proceedings of the 21st International Conference on World Wide Web; Apr 16-20, 2012; Lyon, France. [doi: [10.1145/2187836.2187907](https://doi.org/10.1145/2187836.2187907)]
3. Eysenbach G. Infodemiology: the epidemiology of (mis)information. *Am J Med*. Dec 15, 2002;113(9):763-765. [doi: [10.1016/s0002-9343\(02\)01473-0](https://doi.org/10.1016/s0002-9343(02)01473-0)] [Medline: [12517369](https://pubmed.ncbi.nlm.nih.gov/12517369/)]
4. Qazvinian V, Rosengren E, Radev DR, Barzilay MQ, Johnson R. Rumor has it: identifying misinformation in microblogs. Presented at: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; Jul 27-31, 2011:1589-1599; Edinburgh, Scotland. URL: <https://aclanthology.org/D11-1147/> [Accessed 2025-08-04]
5. Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science*. Mar 9, 2018;359(6380):1146-1151. [doi: [10.1126/science.aap9559](https://doi.org/10.1126/science.aap9559)] [Medline: [29590045](https://pubmed.ncbi.nlm.nih.gov/29590045/)]

6. Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H. The pandemic of social media panic travels faster than the COVID-19 outbreak. *J Travel Med*. May 18, 2020;27(3):taaa031. [doi: [10.1093/jtm/taaa031](https://doi.org/10.1093/jtm/taaa031)] [Medline: [32125413](https://pubmed.ncbi.nlm.nih.gov/32125413/)]
7. Roozenbeek J, Schneider CR, Dryhurst S, et al. Susceptibility to misinformation about COVID-19 around the world. *R Soc Open Sci*. Oct 2020;7(10):201199. [doi: [10.1098/rsos.201199](https://doi.org/10.1098/rsos.201199)] [Medline: [33204475](https://pubmed.ncbi.nlm.nih.gov/33204475/)]
8. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health*. Apr 2, 2020;41:433-451. [doi: [10.1146/annurev-publhealth-040119-094127](https://doi.org/10.1146/annurev-publhealth-040119-094127)] [Medline: [31874069](https://pubmed.ncbi.nlm.nih.gov/31874069/)]
9. Oh SH, Paek HJ, Hove T. Cognitive and emotional dimensions of perceived risk characteristics, genre-specific media effects, and risk perceptions: the case of H1N1 influenza in South Korea. *Asian J Commun*. Jan 2, 2015;25(1):14-32. [doi: [10.1080/01292986.2014.989240](https://doi.org/10.1080/01292986.2014.989240)]
10. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. *J Affect Disord*. Feb 15, 2019;245:270-278. [doi: [10.1016/j.jad.2018.11.037](https://doi.org/10.1016/j.jad.2018.11.037)] [Medline: [30419526](https://pubmed.ncbi.nlm.nih.gov/30419526/)]
11. Fox S, Duggan M. Health Online 2013. Pew Research Center; 2013 Jan 15. URL: <https://www.pewresearch.org/internet/2013/01/15/health-online-2013/> [Accessed 2026-01-26]
12. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. *J Med Internet Res*. May 7, 2015;17(5):e112. [doi: [10.2196/jmir.4018](https://doi.org/10.2196/jmir.4018)] [Medline: [25953147](https://pubmed.ncbi.nlm.nih.gov/25953147/)]
13. Chou WYS, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social media use in the United States: implications for health communication. *J Med Internet Res*. Nov 27, 2009;11(4):e48. [doi: [10.2196/jmir.1249](https://doi.org/10.2196/jmir.1249)] [Medline: [19945947](https://pubmed.ncbi.nlm.nih.gov/19945947/)]
14. Larson HJ. The biggest pandemic risk? Viral misinformation. *Nature*. Oct 2018;562(7727):309. [doi: [10.1038/d41586-018-07034-4](https://doi.org/10.1038/d41586-018-07034-4)] [Medline: [30327527](https://pubmed.ncbi.nlm.nih.gov/30327527/)]
15. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. *J Med Internet Res*. Jan 20, 2021;23(1):e17187. [doi: [10.2196/17187](https://doi.org/10.2196/17187)] [Medline: [33470931](https://pubmed.ncbi.nlm.nih.gov/33470931/)]
16. Islam MS, Sarkar T, Khan SH, et al. COVID-19-related infodemic and its impact on public health: a global social media analysis. *Am J Trop Med Hyg*. Oct 2020;103(4):1621-1629. [doi: [10.4269/ajtmh.20-0812](https://doi.org/10.4269/ajtmh.20-0812)] [Medline: [32783794](https://pubmed.ncbi.nlm.nih.gov/32783794/)]
17. Li Z, Zhang Q, Du X, Ma Y, Wang S. Social media rumor refutation effectiveness: evaluation, modelling and enhancement. *Inf Process Manag*. Jan 2021;58(1):102420. [doi: [10.1016/j.ipm.2020.102420](https://doi.org/10.1016/j.ipm.2020.102420)]
18. Pal A, Chua AYK, Hoe-Lian Goh D. How do users respond to online rumor rebuttals? *Comput Human Behav*. May 2020;106:106243. [doi: [10.1016/j.chb.2019.106243](https://doi.org/10.1016/j.chb.2019.106243)]
19. Maertens R, Roozenbeek J, Basol M, van der Linden S. Long-term effectiveness of inoculation against misinformation: three longitudinal experiments. *J Exp Psychol Appl*. Mar 2021;27(1):1-16. [doi: [10.1037/xap0000315](https://doi.org/10.1037/xap0000315)] [Medline: [33017160](https://pubmed.ncbi.nlm.nih.gov/33017160/)]
20. Chen XK, Na JC, Tan LKW, Chong M, Choy M. Exploring how online responses change in response to debunking messages about COVID-19 on WhatsApp. *Online Inf Rev*. Sep 26, 2022;46(6):1184-1204. [doi: [10.1108/OIR-08-2021-0422](https://doi.org/10.1108/OIR-08-2021-0422)]
21. Walter N, Brooks JJ, Saucier CJ, Suresh S. Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. *Health Commun*. Nov 2021;36(13):1776-1784. [doi: [10.1080/10410236.2020.1794553](https://doi.org/10.1080/10410236.2020.1794553)] [Medline: [32762260](https://pubmed.ncbi.nlm.nih.gov/32762260/)]
22. Li Z, Wu X, Xu L, Liu M, Huang C. Hot topic recognition of health rumors based on anti-rumor articles on the WeChat official account platform: topic modeling. *J Med Internet Res*. Sep 21, 2023;25:e45019. [doi: [10.2196/45019](https://doi.org/10.2196/45019)] [Medline: [37733396](https://pubmed.ncbi.nlm.nih.gov/37733396/)]
23. Xiao Q, Huang W, Zhang X, Wan S, Li X. Internet rumors during the COVID-19 pandemic: dynamics of topics and public psychologies. *Front Public Health*. 2021;9:788848. [doi: [10.3389/fpubh.2021.788848](https://doi.org/10.3389/fpubh.2021.788848)] [Medline: [34988056](https://pubmed.ncbi.nlm.nih.gov/34988056/)]
24. Guess A, Nagler J, Tucker J. Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Sci Adv*. Jan 2019;5(1):eaau4586. [doi: [10.1126/sciadv.aau4586](https://doi.org/10.1126/sciadv.aau4586)] [Medline: [30662946](https://pubmed.ncbi.nlm.nih.gov/30662946/)]
25. Zhao L, Cui H, Qiu X, Wang X, Wang J. SIR rumor spreading model in the new media age. *Phys A Stat Mech Appl*. Feb 2013;392(4):995-1003. [doi: [10.1016/j.physa.2012.09.030](https://doi.org/10.1016/j.physa.2012.09.030)]
26. Castillo C, Mendoza M, Poblete B. Predicting information credibility in time-sensitive social media. *Internet Res*. Oct 14, 2013;23(5):560-588. [doi: [10.1108/IntR-05-2012-0095](https://doi.org/10.1108/IntR-05-2012-0095)]
27. Cook J, Lewandowsky S, Ecker UKH. Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence. *PLoS One*. 2017;12(5):e0175799. [doi: [10.1371/journal.pone.0175799](https://doi.org/10.1371/journal.pone.0175799)] [Medline: [28475576](https://pubmed.ncbi.nlm.nih.gov/28475576/)]

28. Chan MPS, Jones CR, Hall Jamieson K, Albarracín D. Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. *Psychol Sci*. Nov 2017;28(11):1531-1546. [doi: [10.1177/0956797617714579](https://doi.org/10.1177/0956797617714579)] [Medline: [28895452](https://pubmed.ncbi.nlm.nih.gov/28895452/)]
29. Nyhan B, Reifler J. When corrections fail: the persistence of political misperceptions. *Polit Behav*. Jun 2010;32(2):303-330. [doi: [10.1007/s11109-010-9112-2](https://doi.org/10.1007/s11109-010-9112-2)]
30. Yang W, Wang S, Peng Z, Shi C, Ma X, Yang D. Know it to defeat it: exploring health rumor characteristics and debunking efforts on chinese social media during COVID-19 crisis. Presented at: Proceedings of the 16th International AAAI Conference on Web and Social Media; Jun 6-9, 2022:1157-1168; Atlanta, GA. [doi: [10.1609/icwsm.v16i1.19366](https://doi.org/10.1609/icwsm.v16i1.19366)]
31. Sicilia R, Lo Giudice S, Pei Y, Pechenizkiy M, Soda P. Twitter rumour detection in the health domain. *Expert Syst Appl*. Nov 2018;110:33-40. [doi: [10.1016/j.eswa.2018.05.019](https://doi.org/10.1016/j.eswa.2018.05.019)]
32. Budak C, Agrawal D, El Abbadi A. Limiting the spread of misinformation in social networks. Presented at: Proceedings of the 20th International Conference on World Wide Web; Mar 28 to Apr 1, 2011; Hyderabad India. [doi: [10.1145/1963405.1963499](https://doi.org/10.1145/1963405.1963499)]
33. Hassan N, Arslan F, Li C, Tremayne M. Toward automated fact-checking: detecting check-worthy factual claims by claimbuster. Presented at: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2017; Halifax NS, Canada. [doi: [10.1145/3097983.3098131](https://doi.org/10.1145/3097983.3098131)]
34. Li X, Shu C. Research on the influencing factors of new media health rumor recognition ability. Presented at: 2024 11th International Conference on Behavioural and Social Computing (BESC); Aug 16-18, 2024:1-6; Harbin, China. [doi: [10.1109/BESC64747.2024.10780542](https://doi.org/10.1109/BESC64747.2024.10780542)]
35. Chua AYK, Banerjee S. To share or not to share: the role of epistemic belief in online health rumors. *Int J Med Inform*. Dec 2017;108:36-41. [doi: [10.1016/j.ijmedinf.2017.08.010](https://doi.org/10.1016/j.ijmedinf.2017.08.010)] [Medline: [29132629](https://pubmed.ncbi.nlm.nih.gov/29132629/)]
36. Pennycook G, Rand DG. Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*. Jul 2019;188:39-50. [doi: [10.1016/j.cognition.2018.06.011](https://doi.org/10.1016/j.cognition.2018.06.011)] [Medline: [29935897](https://pubmed.ncbi.nlm.nih.gov/29935897/)]
37. Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG. Shifting attention to accuracy can reduce misinformation online. *Nature*. Apr 2021;592(7855):590-595. [doi: [10.1038/s41586-021-03344-2](https://doi.org/10.1038/s41586-021-03344-2)] [Medline: [33731933](https://pubmed.ncbi.nlm.nih.gov/33731933/)]
38. Weibo announces third quarter 2024 unaudited financial results. Weibo. 2024. URL: <http://ir.weibo.com/news-releases/news-release-details/weibo-announces-third-quarter-2024-unaudited-financial-results/> [Accessed 2026-01-26]
39. Vraga EK, Bode L. Using expert sources to correct health misinformation in social media. *Sci Commun*. Oct 2017;39(5):621-645. [doi: [10.1177/1075547017731776](https://doi.org/10.1177/1075547017731776)]
40. van der Meer T, Jin Y. Seeking formula for misinformation treatment in public health crises: the effects of corrective information type and source. *Health Commun*. May 2020;35(5):560-575. [doi: [10.1080/10410236.2019.1573295](https://doi.org/10.1080/10410236.2019.1573295)] [Medline: [30761917](https://pubmed.ncbi.nlm.nih.gov/30761917/)]
41. Wang X, Chao F, Yu G. Evaluating rumor debunking effectiveness during the COVID-19 pandemic crisis: utilizing user stance in comments on Sina Weibo. *Front Public Health*. 2021;9:770111. [doi: [10.3389/fpubh.2021.770111](https://doi.org/10.3389/fpubh.2021.770111)]
42. Silla CN, Freitas AA. A survey of hierarchical classification across different application domains. *Data Min Knowl Disc*. Jan 2011;22(1-2):31-72. [doi: [10.1007/s10618-010-0175-9](https://doi.org/10.1007/s10618-010-0175-9)]
43. Kowsari K, Brown DE, Heidarysafa M, Jafari Meimandi K, Gerber MS, Barnes LE. HDLTex: hierarchical deep learning for text classification. Presented at: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA); Dec 18-21, 2017:364-371; Cancun, Mexico. [doi: [10.1109/ICMLA.2017.0-134](https://doi.org/10.1109/ICMLA.2017.0-134)]
44. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. *Hum Commun Res*. Jul 1, 2004;30(3):411-433. [doi: [10.1111/j.1468-2958.2004.tb00738.x](https://doi.org/10.1111/j.1468-2958.2004.tb00738.x)]
45. Tunstall L, Reimers N, Jo UES, et al. Efficient few-shot learning without prompts. *arXiv*. Preprint posted online on Sep 22, 2022. [doi: [10.48550/arXiv.2209.11055](https://doi.org/10.48550/arXiv.2209.11055)]
46. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. *arXiv*. Preprint posted online on Mar 11, 2022. [doi: [10.48550/arXiv.2203.05794](https://doi.org/10.48550/arXiv.2203.05794)]
47. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Front Sociol*. 2022;7:886498. [doi: [10.3389/fsoc.2022.886498](https://doi.org/10.3389/fsoc.2022.886498)] [Medline: [35602001](https://pubmed.ncbi.nlm.nih.gov/35602001/)]
48. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. *arXiv*. Preprint posted online on Aug 27, 2019. [doi: [10.48550/arXiv.1908.10084](https://doi.org/10.48550/arXiv.1908.10084)]
49. SuperCLUE. URL: <https://www.superclueai.com/homepage> [Accessed 2026-04-22]
50. DeepSeek-AI, A Liu, et al. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv*. Preprint posted online on Dec 2, 2025. [doi: [10.48550/arXiv.2512.02556](https://doi.org/10.48550/arXiv.2512.02556)]

51. Law of the People's Republic of China on protection of minors. The National People's Congress of the People's Republic of China. 2021. URL: http://www.npc.gov.cn/englishnpc/c2759/c23934/202109/t20210914_384808.html [Accessed 2026-01-26]
52. The Standing Committee of the National People's Congress. Law of the People's Republic of China on Protection of the Rights and Interests of the Elderly (2018 amendment) [webpage in Chinese]. The National People's Congress of the People's Republic of China. 2019. URL: http://www.npc.gov.cn/npc/c2/c30834/201905/t20190521_296650.html [Accessed 2026-01-28]
53. National Research Council (US) Committee on Diet and Health. Extent and distribution of chronic disease: an overview. In: Diet and Health: Implications for Reducing Chronic Disease Risk. National Academies Press (US); 1989. URL: <https://www.ncbi.nlm.nih.gov/books/NBK218755> [Accessed 2026-01-15]
54. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist Sci*. 2001;16(3):199-231. [doi: [10.1214/ss/1009213726](https://doi.org/10.1214/ss/1009213726)]
55. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016:785-794; San Francisco, CA. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
56. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *arXiv*. Preprint posted online on May 22, 2017. [doi: [10.48550/arXiv.1705.07874](https://doi.org/10.48550/arXiv.1705.07874)]
57. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. [doi: [10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9)]
58. Romero DM, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. Presented at: WWW '11: Proceedings of the 20th international conference on World wide web; Mar 28 to Apr 1, 2011:695-704; New York, NY. [doi: [10.1145/1963405.1963503](https://doi.org/10.1145/1963405.1963503)]
59. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022:507-520; New Orleans, LA. [doi: [10.52202/068431-0037](https://doi.org/10.52202/068431-0037)]
60. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Presented at: WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; Feb 2-6, 2015; Shanghai China. [doi: [10.1145/2684822.2685324](https://doi.org/10.1145/2684822.2685324)]
61. Dieng AB, Ruiz FJR, Blei DM. Topic modeling in embedding spaces. *Trans Assoc Comput Linguist*. Dec 2020;8:439-453. [doi: [10.1162/tacl_a_00325](https://doi.org/10.1162/tacl_a_00325)]
62. Baumgartner FR, Jones BD. *Agendas and Instability in American Politics*. University of Chicago Press; 2010. [doi: [10.7208/chicago/9780226039534.001.0001](https://doi.org/10.7208/chicago/9780226039534.001.0001)] ISBN: 9780226039497
63. Shih TJ, Wijaya R, Brossard D. Media coverage of public health epidemics: linking framing and issue attention cycle toward an integrated theory of print news coverage of epidemics. *Mass Commun Soc*. Apr 7, 2008;11(2):141-160. [doi: [10.1080/15205430701668121](https://doi.org/10.1080/15205430701668121)]
64. Ball-Rokeach SJ, DeFleur ML. A dependency model of mass-media effects. *Communic Res*. Jan 1976;3(1):3-21. [doi: [10.1177/009365027600300101](https://doi.org/10.1177/009365027600300101)]
65. Stone JA. The sharing of pandemic-related information from U.S. government Twitter accounts. *Am Politics Res*. Sep 2024;52(5):467-483. [doi: [10.1177/1532673X241263088](https://doi.org/10.1177/1532673X241263088)]
66. Kada A, Chouikh A, Mellouli S, Prashad AJ, Straus SE, Fahim C. An exploration of Canadian government officials' COVID-19 messages and the public's reaction using social media data. *PLoS One*. 2022;17(9):e0273153. [doi: [10.1371/journal.pone.0273153](https://doi.org/10.1371/journal.pone.0273153)] [Medline: [36054094](https://pubmed.ncbi.nlm.nih.gov/36054094/)]
67. Ahluwalia SC, Edelen MO, Qureshi N, Etchegaray JM. Trust in experts, not trust in national leadership, leads to greater uptake of recommended actions during the COVID-19 pandemic. *Risk Hazards Crisis Public Policy*. Sep 2021;12(3):283-302. [doi: [10.1002/rhc3.12219](https://doi.org/10.1002/rhc3.12219)] [Medline: [34226844](https://pubmed.ncbi.nlm.nih.gov/34226844/)]
68. Kim NK, Woo B, Yang J. Effect of international organizations' direct engagement with the public: information source credibility and the public's attitudes towards COVID-19-related measures. *Jpn J Polit Sci*. Mar 2026;27(1):19-36. [doi: [10.1017/S1468109925100170](https://doi.org/10.1017/S1468109925100170)]
69. Gollust SE, Nagler RH, Fowler EF. The emergence of COVID-19 in the US: a public health and political communication crisis. *J Health Polit Policy Law*. Dec 1, 2020;45(6):967-981. [doi: [10.1215/03616878-8641506](https://doi.org/10.1215/03616878-8641506)] [Medline: [32464658](https://pubmed.ncbi.nlm.nih.gov/32464658/)]
70. Stifjell K, Sandanger TM, Wien C. Exploring online health information-seeking behavior among young adults: scoping review. *J Med Internet Res*. Sep 9, 2025;27:e70379. [doi: [10.2196/70379](https://doi.org/10.2196/70379)] [Medline: [40925001](https://pubmed.ncbi.nlm.nih.gov/40925001/)]
71. Maslow AH. A theory of human motivation. *Psychol Rev*. 1943;50(4):370-396. [doi: [10.1037/h0054346](https://doi.org/10.1037/h0054346)]
72. Conrad P. *The Medicalization of Society: On the Transformation of Human Conditions into Treatable Disorders*. Johns Hopkins University Press; 2007. [Accessed 2026-01-26] ISBN: 9780801885853

73. Fredrickson BL, Roberts TA. Objectification theory: toward understanding women's lived experiences and mental health risks. *Psychol Women Q*. Jun 1997;21(2):173-206. [doi: [10.1111/j.1471-6402.1997.tb00108.x](https://doi.org/10.1111/j.1471-6402.1997.tb00108.x)]
74. Ek S. Gender differences in health information behaviour: a Finnish population-based survey. *Health Promot Int*. Sep 2015;30(3):736-745. [doi: [10.1093/heapro/dat063](https://doi.org/10.1093/heapro/dat063)] [Medline: [23985248](https://pubmed.ncbi.nlm.nih.gov/23985248/)]
75. Wang MP, Viswanath K, Lam TH, Wang X, Chan SS. Social determinants of health information seeking among Chinese adults in Hong Kong. *PLoS One*. 2013;8(8):e73049. [doi: [10.1371/journal.pone.0073049](https://doi.org/10.1371/journal.pone.0073049)] [Medline: [24009729](https://pubmed.ncbi.nlm.nih.gov/24009729/)]
76. Zhu C, Zeng R, Zhang W, Evans R, He R. Pregnancy-related information seeking and sharing in the social media era among expectant mothers: qualitative study. *J Med Internet Res*. Dec 4, 2019;21(12):e13694. [doi: [10.2196/13694](https://doi.org/10.2196/13694)] [Medline: [31799939](https://pubmed.ncbi.nlm.nih.gov/31799939/)]
77. Connell R. *Masculinities*. Routledge; 2020. ISBN: 9781003116479
78. Courtenay WH. Constructions of masculinity and their influence on men's well-being: a theory of gender and health. *Soc Sci Med*. May 2000;50(10):1385-1401. [doi: [10.1016/s0277-9536\(99\)00390-1](https://doi.org/10.1016/s0277-9536(99)00390-1)] [Medline: [10741575](https://pubmed.ncbi.nlm.nih.gov/10741575/)]
79. Bucher T. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*. Nov 2012;14(7):1164-1180. [doi: [10.1177/1461444812440159](https://doi.org/10.1177/1461444812440159)]
80. Dey M. Weibo statistics by revenue, market cap, users and facts. *Electro IQ*. 2025. URL: <https://electroiq.com/stats/weibo-statistics/> [Accessed 2026-01-26]
81. Cutrona SL, Mazor KM, Vieux SN, Luger TM, Volkman JE, Finney Rutten LJ. Health information-seeking on behalf of others: characteristics of "surrogate seekers". *J Cancer Educ*. Mar 2015;30(1):12-19. [doi: [10.1007/s13187-014-0701-3](https://doi.org/10.1007/s13187-014-0701-3)] [Medline: [24989816](https://pubmed.ncbi.nlm.nih.gov/24989816/)]
82. Petty RE, Cacioppo JT. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag; 1986. URL: <https://link.springer.com/book/10.1007/978-1-4612-4964-1> [Accessed 2026-01-26] ISBN: 9781461249641
83. Blau P. *Exchange and Power in Social Life*. Routledge; 2017. [Accessed 2026-01-28] ISBN: 9780887386282
84. Surma J. Social exchange in online social networks. The reciprocity phenomenon on Facebook. *Comput Commun*. Jan 2016;73:342-346. [doi: [10.1016/j.comcom.2015.06.017](https://doi.org/10.1016/j.comcom.2015.06.017)]
85. Merton RK. The Matthew effect in science. The reward and communication systems of science are considered. *Science*. Jan 5, 1968;159(3810):56-63. [doi: [10.1126/science.159.3810.56](https://doi.org/10.1126/science.159.3810.56)] [Medline: [5634379](https://pubmed.ncbi.nlm.nih.gov/5634379/)]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
IAA: interannotator agreement
LLM: large language model
PPE: personal protective equipment
QWK: Quadratically Weighted Cohen κ
RMSE: root-mean-squared error
RQ: research question
SetFit: Sentence Transformer Fine-tuning
SHAP: Shapley Additive Explanations
XGBoost: Extreme Gradient Boosting

Edited by Amaryllis Mavragani; peer-reviewed by Babatunde Balogun, Nikolaos Yiannakoulis, Rui Guo; submitted 29.Jan.2026; final revised version received 04.Jul.2026; accepted 15.Jul.2026; published 10.Sep.2026

Please cite as:

Fang Y, He C, Hu Y, Tian X
Health-Related Rumor Debunking on Sina Weibo in China (2009-2024): 15-Year Retrospective Infodemiology Study
J Med Internet Res 2026;28:e91516
 URL: <https://www.jmir.org/2026/1/e91516>
 doi: [10.2196/91516](https://doi.org/10.2196/91516)

© Yuan Fang, Chengwu He, Yejinxuan Hu, Xianyun Tian. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 10.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871),

is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.