How Do Social Bots Participate in Misinformation Spread? A Comprehensive Dataset and Analysis

Herun Wan[1, 2] **Minnan Luo**[1, 2, *] **Zihan Ma**[1, 3]
Guang Dai[4] **Xiang Zhao**[5]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[3]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, China
[4]SGIT AI Lab, State Grid Corporation of China
[5]National University of Defense Technology

wanherun@stu.xjtu.edu.cn minnluo@xjtu.edu.cn

https://whr000001.github.io/MisBot/

Abstract

Social media platforms provide an ideal environment to spread misinformation, where social bots can accelerate the spread. This paper explores the interplay between social bots and misinformation on the Sina Weibo platform. We construct a large-scale dataset that includes annotations for both misinformation and social bots. From the misinformation perspective, the dataset is multimodal, containing 11,393 pieces of misinformation and 16,416 pieces of verified information. From the social bot perspective, this dataset contains 65,749 social bots and 345,886 genuine accounts, annotated using a weakly supervised annotator. Extensive experiments demonstrate the comprehensiveness of the dataset, the clear distinction between misinformation and real information, and the high quality of social bot annotations. Further analysis illustrates that: (i) social bots are deeply involved in information spread; (ii) misinformation with the same topics has similar content, providing the basis of echo chambers, and social bots would amplify this phenomenon; and (iii) social bots generate similar content aiming to manipulate public opinions.

1 Introduction

Social media platforms, like 𝕏 (Twitter) and Weibo, have become major information sources, and information spreads faster than traditional media. Due to the nature of such platforms, there have been attempts to disseminate misinformation, which could polarize society (Azzimonti and Fernandes, 2023) and impact the economy (Zhou et al., 2024). Meanwhile, besides attracting genuine users, the social platform also becomes an ideal breeding ground for malicious social bots (Cresci, 2020) due to the

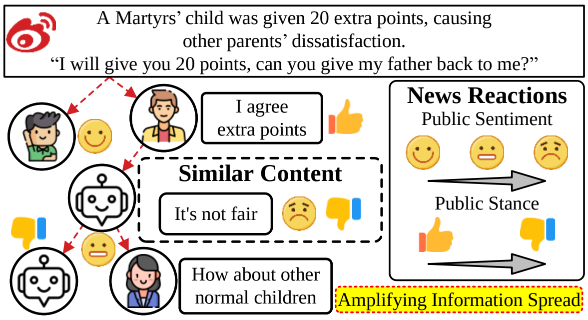


Figure 1: An example of social bots participating in information spread. Social bots would publish similar content to manipulate public sentiment and stance, leading to a shift in public opinion.

straightforward operation. Social bots are proven behind many online perils, including election interference (Ng et al., 2022) and hate speech propaganda (Stella et al., 2019). Social bots are natural message amplifiers (Caldarelli et al., 2020), increasing the risk of spreading misinformation (Huang et al., 2022). Namely, misinformation and social bots are two major factors harming online security. They might work together to amplify negative impact, where Figure 1 presents an example.

Researchers make efforts to fight the never-ending plague of misinformation and malicious social bots. They mainly propose automatic detectors to identify misinformation (Shu et al., 2019) and social bots (Yang et al., 2022). Meanwhile, researchers also explore how different types of content (Nan et al., 2021) or propagation patterns (Vosoughi et al., 2018) influence misinformation spread. From the social bot perspective, bot communities (Tan et al., 2023b) and bots' repost behaviors (Elmas et al., 2022) have been investigated. While many works have provided valuable insights into investigating misinformation and so-

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31493–31516

cial bots, relatively little attention (Wang et al., 2018; Himelein-Wachowiak et al., 2021) has been paid to the interplay between them.

This paper aims to bridge the gap of existing works, exploring the interplay between misinformation and social bots. We propose MisBot[1], a dataset which simultaneously contains information and annotations of misinformation and social bots (§2). Specifically, we first define the structure of MisBot. We then collect misinformation from Weibo's official management center[2]. After that, we collect real information from two credible sources to ensure the authenticity. We finally propose a weakly supervised annotator to label the users involved in the dissemination of information. From the misinformation perspective, MisBot contains multiple modalities, including post content, comments, repost messages, images, and videos. MisBot includes 11,393 misinformation instances and 16,416 real information instances. From the social bot perspective, MisBot includes 952,955 users participating in the information spread, covering 65,749 annotated social bots and 345,886 genuine accounts. Extensive experiments (§3) prove that (i) MisBot is the most comprehensive and the only one with misinformation and social bot annotations, (ii) misinformation and real information are distinguishable, where a simple detector achieves 95.2% accuracy, and (iii) MisBot has high social bot annotation quality, where human evaluations prove it. Further analysis illustrates (§4) that (i) social bots are deeply involved in information spread, where 29.3% users who repost misinformation are social bots; (ii) misinformation with the same topics has similar content, providing the basis of echo chambers, and social bots amplify this phenomenon; and (iii) social bots generate similar content aiming to manipulate public opinions, including sentiments and stances.

2 MisBot Dataset

The collection process of MisBot consists of four components: (i) **Data Structure** defines the dataset structure; (ii) **Misinformation Collection** collects multiple modalities in misinformation; (iii) **Real Information Collection** collects real information from two sources; and (iv) **Weakly Supervised User Annotation** trains a weakly supervised annotator to automatically annotate accounts.

2.1 Data Structure

Users publish posts to spread information on the Weibo platform, thus, we annotate user posts as misinformation or real information. From this perspective, each instance is represented as $A = \{s, \mathcal{G}_{repost}, \mathcal{G}_{comment}, I, V, y\}$. It contains textual content s, repost graph $\mathcal{G}_{repost}$, comment graphs $\mathcal{G}_{comment}$, images I, videos V, and corresponding label y. From the account perspective, each instance is represented as $U = \{F, T, y\}$. It contains the attribute set F, published posts T, and the corresponding label y. Meanwhile, we believe a user participates in the spread of a post if this user reposts, comments, or likes this post. Some cases in MisBot are provided in Appendix A.1.

2.2 Misinformation Collection

We collect posts flagged as misinformation from Weibo's official management center, where we provide the platform overview in Appendix A.2 for readers who cannot log in. This platform presents posts containing misinformation judged by platform moderators or police. Besides, it provides a brief **judgment** to explain why the post is flagged as misinformation, which provides a basis for identifying topics of misinformation. An example is provided in Appendix A.3. We have collected all the misinformation since this platform was established. Specifically, the misinformation collected was published between April 2018 and April 2024. We spent about 10 months collecting 11,393 pieces of misinformation.

2.3 Real Information Collection

Existing misinformation datasets generally suffer from potential data bias (Chen et al., 2023), especially entity biases (Zhu et al., 2022). It means that the entity distributions in misinformation and real information differ, influencing models' generalization ability to unseen data. Thus, we design an entity debiasing method to mitigate entity biases. We first employ a keyphrase extractor (Liang et al., 2021) to obtain key entities from each misinformation. After filtering uncommon entities, we get 1,961 entities, where we present the filter rules in Appendix A.4. We finally query the key entities using the Weibo search engine in trusted sources to get real information. An overview of the search engine is provided in Appendix A.5 for readers who

[1]The main language of MisBot is Chinese.

[2]https://service.account.weibo.com/, being available for users who have logged in.

cannot log in. To ensure the authenticity and diversity of real information, we collect real information from two sources:

- **Verified news media** is an official news account certified by the Weibo platform, which contains a red "verified" symbol and a verified reason, where we provide the statistics of the verified news accounts in Appendix A.6.

- **Trends on the platform** contains posts sparking a lot of discussion in a short period.

Due to the moderation of Weibo, we assume these two sources are truthful, where we discuss it further in Appendix A.7 and quantitatively prove it in §3. We obtained 8,317 and 8,099 pieces of real information, respectively.

2.4 Weakly Supervised User Annotation

Manual annotation or crowd-sourcing is labor-intensive and not feasible with large-scale datasets. Meanwhile, to ensure the scalability of MISBOT, we propose a weakly supervised learning strategy, enabling automatic annotation. The construction of the weakly supervised annotator contains (i) preprocessing, (ii) training, and (iii) inference phases.

Preprocessing Phase This phase aims to collect the training dataset for the weakly supervised annotator. We first collected 100,000 random accounts. Due to the randomness, these accounts could represent the entire Weibo environment, ensuring the diversity of accounts. We employ crowd-sourcing to annotate them, where the human annotators are familiar with social media. Following existing works (Feng et al., 2021b, 2022), we summarize a brief criteria for identifying a social bot on Weibo and write a guideline document for human annotators, where we provide the document in Appendix A.8. Notably, social bot annotation is subjective, where the average Fleiss' Kappa is 0.4281 as shown in §3. Thus, we do not directly define what a social bot is, but only provide a brief guideline document and cases. Inspired by existing work (Feng et al., 2021b), we determine 20 standard accounts that are easy to identify. Each annotator should also annotate 20 standard accounts, and annotators who achieve more than 80% accuracy on standard accounts are reliable. We ensure that each account is annotated by three reliable human annotators. We totally recruited 315 annotators and spent 60,000 yuan and 60 days, where we provide details in Appendix A.9. We employ major voting to obtain the final annotations in this phase.

Based on human annotators' feedback, we filter in active accounts in MISBOT, where we provide the filter rules in Appendix A.10. We focus on active accounts for three reasons:

- We aim to explore the involvement of social bots in misinformation and real information spread, where inactive users hardly participate in information spread.

- Annotators mainly rely on posts in users' timelines to make judgments, whereas inactive accounts cannot provide enough information to obtain reliable annotations.

- Mainstream social bot detectors analyze accounts' posts to identify bots, and we follow this to construct an annotation model. We employ active users to ensure credibility.

We obtained 48,536 active accounts from the 100,000 accounts, of which 18,132 are social bots and 30,404 are genuine accounts.

Training Phase Different machine bot detectors have their strengths and weaknesses in the face of multiple social bots (Sayyadiharikandeh et al., 2020). Thus, we propose to employ multiple detectors as experts and employ an ensemble strategy to obtain the final annotations. In this phase, we leverage the following detectors:

- **Feature-based detectors** leverage feature engineering on user attributes and adopt classic machine learning algorithms to identify social bots. We employ various attributes: (i) **numerical**: *follower count*, *following count*, and *status count*; and (ii) **categorical**: *verified*, *svip*, *account type*, and *svip level*. We employ MLP layers, random forests, and Adaboost as detectors.

- **Content-based detectors** encode user-generated textual content, where we employ *name*, *description*, and *posts*. We employ encoder-based language models, including BERT (Devlin et al., 2019) and DeBERTa (He et al., 2021) to obtain textual representations and employ an MLP layer to identify social bots.

- **Ensemble detectors** concatenate the attribute and textual representations and employ an MLP layer to identify social bots.

The descriptions and settings of experts are provided in Appendix A.11. We create an 8:1:1 split

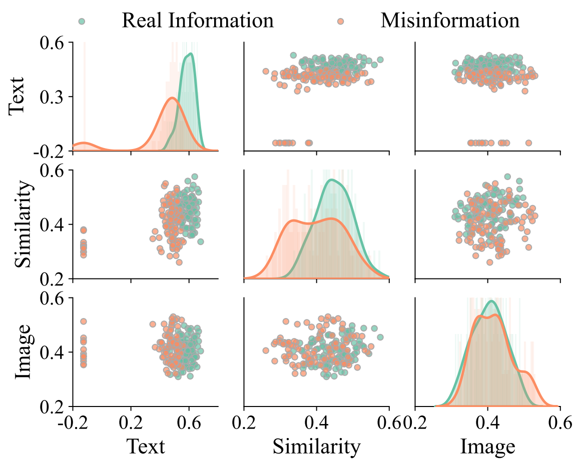


Figure 2: The joint distributions of three content consistency metrics for misinformation and real information. Misinformation and real information illustrate different distributions, especially in **Text** and **Similarity**.

for the users from the preprocessing phase as train, validation, and test sets to train each expert.

Inference Phase This phase annotates accounts based on the predictions from multiple experts. To ensure annotation quality, we filter in the experts achieving 80% accuracy, which is the same standard as human annotators, on the validation set. After that, a conventional method to integrate multiple predictions is to employ majority voting or train an MLP classifier on the validation set (Bach et al., 2017; Feng et al., 2022). Since the likelihood from classifiers may not accurately reflect true probabilities (Guo et al., 2017), also known as *miscalibrated*, we calibrate the likelihoods before the ensemble. We employ temperature scaling (Guo et al., 2017) and select the best temperature on the validation set, where we provide the temperature settings in Appendix A.12. We finally average the calibrated likelihoods to obtain the final annotations. Among the 952,955 accounts that participate in information spread in MisBot, 411,635 are active, of which 65,749 are social bots and 345,886 are genuine accounts.

3 Basic Analysis of MisBot

MisBot is the most comprehensive. We compare MisBot with the recent datasets for misinformation and social bots, illustrated in Table 1. MisBot is the only dataset simultaneously containing misinformation and social bot annotations. Meanwhile, from the misinformation perspective, MisBot contains the most complete multi-modal information, including textual content, user comments, repost messages, images, videos, and related users. MisBot is the largest and contains the richest visual modal data for misinformation.

Misinformation and real information in MisBot are distinguishable. We aim to explore the role of social bots in amplifying misinformation spread, which requires misinformation and real information to be distinguishable. Thus, we analyze whether misinformation and real information are distinguishable from two perspectives: *data distribution* and *misinformation detector*.

From the *data distribution* perspective, we first explore the differences in content consistency between misinformation and real information. We employ three metrics: (i) **Text** to evaluate the text consistency of a specific instance and all instances; (ii) **Image** to evaluate the image consistency of a specific instance and all instances; and (iii) **Similarity** to evaluate the consistency of text and image in a specific instance. We provide the calculation formula in Appendix B.1 and present the joint distributions in Figure 2. It illustrates that misinformation and real information present distinct consistency. Specifically, real information presents higher **Text** and **Similarity**. Namely, we could conclude that misinformation and real information are distinguishable in terms of consistency.

To further capture the image differences between misinformation and real information, we present the distribution of image categories and sentiments in Figure 15 in Appendix B.2. It illustrates that misinformation and real information present distinct distributions. Specifically, real information would contain more neutral images while misinformation would contain more screenshots.

From the *misinformation detector* perspective, we design a simple misinformation detector to verify whether the detector could identify misinformation in MisBot, where we provide the details of this model in Appendix B.3. We present the performance of the detector and ablation variants in Table 2. This simple detector achieves remarkable performance, where the accuracy reaches 95.2%. The ideal performance proves that misinformation and real information are easily distinguished by a machine detector, which helps explore the differences between social bots in spreading misinformation and real information. Meanwhile, the detector without *interaction* drops to 77.3% on f1-score, illustrating the effectiveness of user reac-

Dataset	Modalities						Statistics					
	Content	Comment	Repost	Image	Video	User	Post	Image	Video	User	Bots	Human
Datasets for misinformation detection.												
(Shu et al., 2020)*	✓	✓	✓	✓		✓	23,196	19,200	0	**2,063,442**	0	0
(Nan et al., 2021)*	✓						9,128	0	0	0	0	0
(Li et al., 2022)					✓		700	0	700	0	0	0
(Qi et al., 2023)*		✓			✓	✓	3,654	0	3,654	3,654	0	0
(Hu et al., 2023)	✓	✓		✓	✓		14,700	14,700	0	0	0	0
(Li et al., 2024)*	✓	✓		✓		✓	23,789	10,178	0	803,779	0	0
Datasets for social bot detection.												
(Feng et al., 2021b)						✓	0	0	0	229,580	6,589	5,237
(Feng et al., 2022)		✓	✓			✓	0	0	0	1,000,000	**139,943**	**860,057**
(Shi et al., 2023)*		✓	✓			✓	0	0	0	410,199	2,748	7,451
Datasets for the interplay between misinformation and social bots.												
MisBot	✓	✓	✓	✓	✓	✓	**27,809**	**61,714**	**7,328**	952,955	65,749	345,886

Table 1: Summary of our dataset and recent datasets for misinformation and social bots. We first check each dataset's modality and then report the related statistics. The $\star$ denotes that the publisher does not provide the original data in the corresponding paper. Our dataset is the largest and the only one with misinformation and social bot annotations, containing 27,809 instances.

Models	Accuracy	F1-score	Precision	Recall
Vanilla	$95.2_{\pm 0.6}$	$92.3_{\pm 0.8}$	$93.7_{\pm 1.7}$	$91.0_{\pm 1.0}$
w/o *Interaction*	$81.6^{\star}_{\pm 4.5}$ 14.2%↓	$77.3^{\star}_{\pm 4.1}$ 16.2%↓	$64.4^{\star}_{\pm 5.9}$ 31.3%↓	$97.3^{\star}_{\pm 1.2}$ 7.0%↑
w/o *Vision*	$94.1^{\star}_{\pm 0.5}$ 1.1%↓	$90.3^{\star}_{\pm 1.0}$ 2.2%↓	$94.2^{\dagger}_{\pm 1.2}$ 0.4%↑	$86.8^{\star}_{\pm 2.1}$ 4.6%↓
w/o *Extra*	$78.5^{\star}_{\pm 5.2}$ 17.5%↓	$74.5^{\star}_{\pm 4.3}$ 19.3%↓	$60.6^{\star}_{\pm 6.0}$ 35.4%↓	$97.3^{\star}_{\pm 1.0}$ 6.9%↑

Table 2: Performance of baseline and variants. We report the mean and standard deviation of ten-fold cross-validation. We also report the performance changes and conduct the paired t-test with vanilla, where $\star$ denotes the p-value is less than 0.0005 and $\dagger$ denotes otherwise. Misinformation and real information are distinguishable with the help of user interactions.

tions, which coincides with our speculation that social bots might have different social patterns in misinformation and real information. We also provide a complete analysis of the ablation study in Appendix B.4.

MisBot has high social bot annotation quality, where the weakly supervised annotator is reliable. The construction of the weakly supervised annotator contains three phases, where we have proven that each phase is reliable:

- **Preprocessing phase.** We recruited 315 human annotators, each of whom annotated 1,000 accounts and 20 standard accounts (the annotators did not know the standard accounts). Among them, 300 human annotators achieved more than 80% accuracy on the standard accounts. The average accuracy of the reliable annotators on the standard accounts is 93.75%. For the agreement between human annotators, the average Fleiss' Kappa is 0.4281, showing moderate agreement.

- **Training phase.** We employed multiple detectors aiming to identify various social bots. To ensure the annotator's credibility, we filtered in detectors achieving 80% accuracy and obtained 4 detectors. The accuracy on the test set reaches 85.03%, which is higher than TwiBot-20 (Feng et al., 2021b) and TwiBot-22 (Feng et al., 2022), illustrating credibility. We also provide the performance of each detector and the corresponding temperature in Appendix B.5.

- **Inference phase.** We randomly sample 50 social bots and 50 genuine accounts in MisBot and manually annotate them through a human expert. The Cohen's Kappa between the human expert and the automatic annotator is 0.74, showing good agreement.

4 Misinformation and Social Bots

Social bots are deeply involved in information spread. We first check the bot percentage:

- The whole MisBot contains 952,955 accounts, of which 411,635 are active. There are 65,749 social bots, accounting for 15.97%.

- Among 5,750 accounts publishing misinformation, there are 3,799 active accounts. There are 767 social bots, accounting for 20.19%.

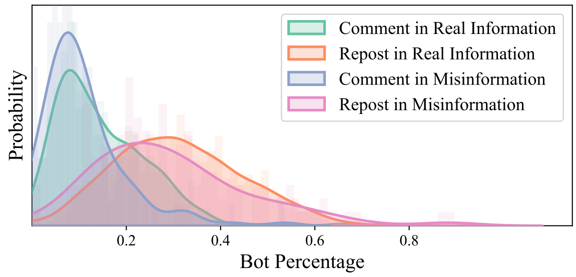


Figure 3: Probability density distributions of the percentage of social bots in information reposting and commenting. Social bots are deeply involved in information reposting and commenting.

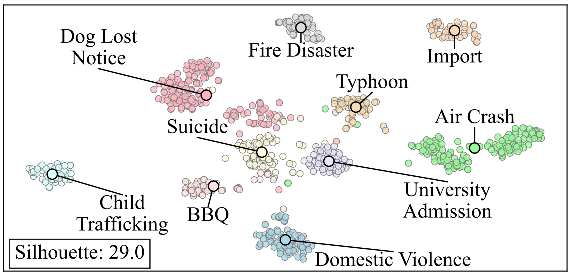


Figure 4: Visualization of misinformation content representations within the largest 10 clusters. Each dot corresponds to a misinformation instance colored according to its topic. The topic labels annotated by the **judgment** are plotted at each cluster center. We also calculate the silhouette score ($\times 100$). The cohesive clusters indicate misinformation about the same topic having similar content, providing the basis of echo chambers.

- Among 226,235 accounts participating in the misinformation spread, 95,360 are active. There are 13,020 social bots, accounting for 13.65%.
- Among 749,763 accounts participating in the real information spread, 325,414 are active. There are 54,253 social bots, accounting for 16.67%.

Figure 3 further presents the distribution of social bots in information reposting and commenting. The average bot percentage of misinformation reposting and commenting is 29.3% and 10.9%, respectively, while the percentage of real information reposting and commenting is 31.1% and 14.7%. It illustrates that the distribution of misinformation and real information is similar, with slightly more social bots participating in spreading real information than misinformation. Meanwhile, reposting tends to have a higher bot percentage than commenting. Thus, we could conclude that social bots are deeply involved in information spread, where the main spread method is to repost information.

Misinformation with the same topics has similar content, providing the basis of echo chambers, and social bots amplify this phenomenon. We first group all pieces of misinformation into clusters with the same topics according to the **judgment**, where we provide the clustering algorithm in Appendix C.1. We group 11,393 pieces of misinformation into 2,270 clusters, each of which represents a specific topic or event, *e.g.*, "The last two minutes of the air crash". We aim to explore the textual content similarity of misinformation with the same topics and across different topics.

We first select the 10 largest clusters as representatives, since there is a long-tail effect in cluster size, where we present the selected clusters in Appendix C.2. We visualize the misinformation content representations in Figure 4, which shows the BERT representation using t-SNE dimensionality reduction. It illustrates that the clusters are cohesive, where the silhouette score is 0.29. Namely, each cluster shares similar content while different clusters share significant differences. It suggests that the misinformation environment is homogeneous, providing the basis for echo chambers.

We conduct further quantitative analysis by calculating *semantic-level* and *token-level* pairwise scores between two instances, where higher scores mean the content of the two instances is more similar. For *semantic level*, we employ the cosine similarity of the BERT representations, while for *token level*, we leverage the ROUGE-L score, where we provide the detailed calculation in Appendix C.3. For *semantic level*, the average value within the same cluster is 0.9448, and the others' average is 0.5847. For *token level*, the average value within the same cluster is 0.7815, and the others' average is 0.0773. We also present completed values in Figure 17 in Appendix C.4. The quantitative results emphasize that misinformation with the same topics has similar content, and misinformation with different topics has distinct content.

We finally explore the patterns of social bots in misinformation. We consider an account a potential *echo chamber member* if it participates in at least two misinformation discussions (repost, comment, or like) in the same cluster. Figure 5 presents the distribution of bot percentage among echo chamber members and non-members within various clusters. It illustrates that around 18% non-members are social bots. Meanwhile, the members do not contain bots in about half of the clusters. However,

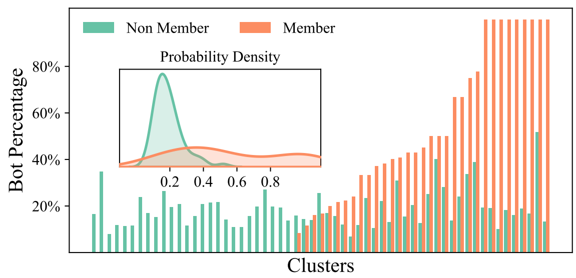


Figure 5: Bot percentage distribution comparison between echo chamber members and non-members across various clusters. Bot percentage among echo chamber members is generally higher than among non-members.

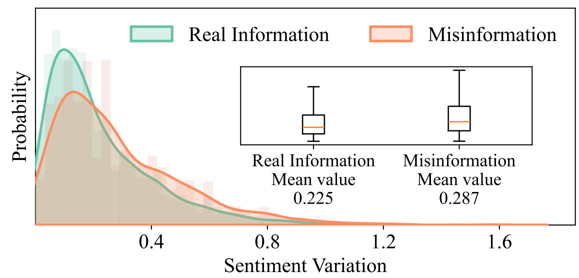


Figure 6: The distribution of neutral sentiment variation. Public sentiment changes are dramatic during information spread, with misinformation slightly more drastic.

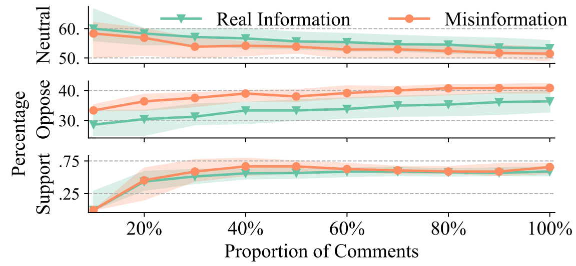


Figure 7: The proportion of comments with different stances as the comments increases. Public stances become increasingly polarized, where misinformation contains more comments with clear stances.

in the other half, members exhibit a higher bot percentage across most clusters compared to non-members, reaching up to 50% in many clusters. We speculate that social bots engage in discussions involving misinformation on specific topics, thereby reinforcing the echo chamber effect.

Social bots generate similar content, aiming to manipulate public opinions. Online information consumers are reluctant to process information deliberately (Möller et al., 2020), becoming susceptible to cognitive biases (Pennycook et al., 2018; Vosoughi et al., 2018). We aim to explore how public opinion changes and how social bots potentially manipulate it. We focus on how public sentiments and stances change in MISBOT. We employ two existing classifiers to obtain the sentiments and stances since it is not our contribution, where we provide the details in Appendix C.5.

- For *public sentiments*, we categorize sentiments into *neutral* and *non-neutral* (including *happy*, *angry*, *surprised*, *sad*, and *fearful*). Figure 18 in Appendix C.6 presents sentiment distribution in different social texts. It illustrates that misinformation would publish more emotional content while real information would naturally be reported. On the other hand, public reactions are always emotional, where misinformation shows more anger while real information shows more happiness. Thus, public sentiments are emotional. We further explore the degree or extent to which public sentiments change over the information spread, introducing a variation measure:

$$v_\Delta = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|,$$

where $f(x_k)$ denotes neutral sentiment proportion at time x_k and we provide the details of x_k in Appendix C.7. Figure 6 visualizes sentiment variation distribution, where a larger value means a more drastic change. The average values of misinformation and real information reach 0.287 and 0.225. It illustrates that public sentiment changes are dramatic during information spread.

- For *public stances*, we categorize stances into *support*, *oppose*, and *neutral*. Figure 7 presents the proportion of each stance with the comments increasing over time. A striking finding is that only about 1% accounts explicitly expressed a supportive stance, while the majority are neutral or opposed. Meanwhile, misinformation consistently presents higher opposition and lower neutrality. It illustrates that public stances become more polarized as the information spreads, where the neutral ratio suffers a drop of around 11%.

Therefore, we can conclude that as the information spreads, public opinions, including sentiment and stance, become polarized, especially regarding misinformation. We then quantitatively prove the correlation between polarization and social bots by the Pearson correlation coefficient: the number of social bots demonstrates strong correlations with the number of comments with non-neutral stances

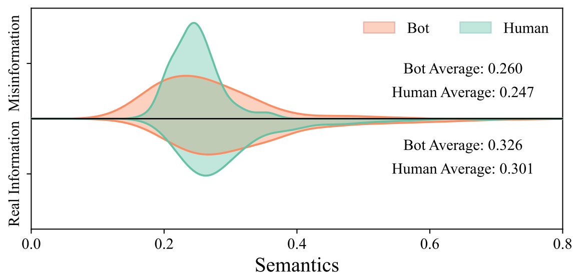


Figure 8: Distribution of social bots' and humans' semantic similarities, where social bots present higher similarities. Namely, social bots would publish similar content to manipulate public opinion.

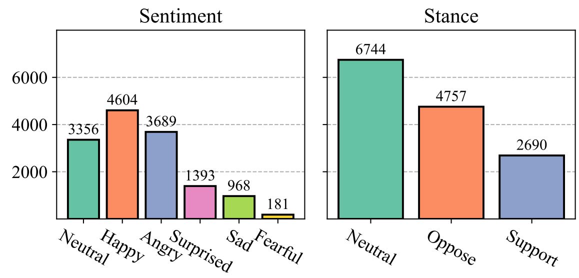


Figure 9: The sentiments and stances of comments published by social bots. Social bots would publish polarized content, manipulating public opinions.

(r = 0.6661) and sentiments (r = 0.6750). We also provide the completed coefficient in Appendix C.8. The relatively high correlation coefficients indicate that social bots might influence public opinion.

We further explore social bot characteristics in information spread. We first calculate the semantic similarity of a specific account, where a higher value means that this account would publish more similar content. We present the detailed calculation method in Appendix C.9 and present the results in Figure 8. It illustrates that social bots generally present higher values than humans. Social bots would publish similar content to amplify the *bandwagon effect*, where online users adopt behaviors or actions simply because others are doing so, influencing the information spread (Wang and Zhu, 2019). We then identify the sentiments and stances of social texts generated by social bots and present the results in Figure 9. It demonstrates that social bots publish more emotional content and comments with clear stances. The results enhance the finding that social bots generate similar content, aiming to manipulate public opinions.

5 Related Work

5.1 Misinformation Detection

Mainstream detectors focuses on the information content, including text (Hartl and Kruschwitz, 2022; Xiao et al., 2024; Gong et al., 2024), images (Liu et al., 2023a; Zhang et al., 2024b,d), and videos (Tan et al., 2023a; Bu et al., 2024; Zeng et al., 2025). They extract features such as emotion (Zhang et al., 2021; Wan et al., 2025c) and employ neural networks such as graph neural networks (Tao et al., 2024; Zhang et al., 2024f; Lu et al., 2024; Wan et al., 2024b) or neurosymbolic reasoning (Dong et al., 2024) to characterize information. Besides information content, the context, such as user interactions (Shu et al., 2019; Lu and Li, 2020), user profile (Sun et al., 2023; Xu et al., 2024), and evidence (Chen et al., 2024; Wan et al., 2025a) provides helpful signals to detect misinformation. They would model propagation patterns (Cui and Jia, 2024), construct news environments (Yin et al., 2024), or extract multi-hop facts (Zhang et al., 2024a) to enhance detection performance. Recently, to combat LLM-generated misinformation (Zhang et al., 2024e; Venkatraman et al., 2024; Wan et al., 2025b), models employing LLMs (Wan et al., 2024a; Nan et al., 2024) through prompting (Guan et al., 2024; Hu et al., 2024) and in-context learning (Wang et al., 2024) have been proposed.

5.2 Social Bot Detection

Social bot detectors fall into feature-, content-, and graph-based. Feature-based models conduct feature engineering for accounts (Feng et al., 2021a; Hays et al., 2023). Content-based models employ NLP techniques (Lei et al., 2023; Cai et al., 2024) to characterize the content. Graph-based models model user interactions as graph structures and employ graph neural networks (Feng et al., 2021c; Yang et al., 2023b; Zhou et al., 2023; Liu et al., 2024) in a semi-supervised way to identify bots. Many researchers are committed to exploring the risks and opportunities LLMs bring to bot detection (Tan and Jiang, 2023; Feng et al., 2024).

5.3 Social Media Safety

Social media safety has become more crucial (Mou et al., 2024), where misinformation and social bots are two main factors harming online safety. Numerous datasets for misinformation (Li et al., 2024; Qazi et al., 2024; Lin et al., 2024; Chen and Shu, 2024) and social bots (Feng et al., 2021b, 2022; Shi et al., 2023) are proposed. Based on these datasets, the generalization of detectors (Zhang et al., 2024c; Assenmacher et al., 2024), misinformation propagation pattern (Aghajari, 2023; Ashkinaze et al., 2024), how to mitigate misinformation spread (Konstantinou and Karapanos, 2023; Su et al., 2024; Ghosh et al., 2024), health-related misinformation (Yang et al., 2024; Shang et al., 2024), source credibility (Carragher et al., 2024; Mehta and Goldwasser, 2024), user profiling (Morales et al., 2023; Zeng et al., 2024), and bot communities (Liu et al., 2023b; Tan et al., 2023b) are investigated. However, relatively little attention has been paid to the interplay between misinformation and social bots, thus, we bridge the gap in this paper.

6 Conclusion

In this paper, we proposed a novel dataset named MISBOT containing information and annotations of misinformation and social bots. MISBOT is the most comprehensive; misinformation and real information are distinguishable; and social bots have high annotation quality. Extensive analysis illustrates that (i) social bots are deeply involved in information spread; (ii) misinformation provides the basis of echo chambers, and social bots amplify this phenomenon; (iii) social bots generate similar content aiming to manipulate public opinions.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 62192781, No. 62272374), the Natural Science Foundation of Shaanxi Province (No. 2024JC-JCQN-62), the Key Research and Development Project in Shaanxi Province (No. 2023GXLH-024), the Project of China Knowledge Center for Engineering Science and Technology, the Project of Chinese Academy of Engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China", and the K. C. Wong Education Foundation. This work was completed during the internship at SGIT AI Lab, State Grid Corporation of China.

Limitation

Our proposed dataset is the largest containing misinformation and social bot annotations simultaneously. Meanwhile, it contains multiple modalities, including images and videos, and user interactions. However, due to the focus on news spread, it does not contain interactions like the friend relationship, missing potential relations between social bots and genuine accounts. Meanwhile, we propose a weakly supervised framework for annotating social bots, whose accuracy is comparable to that of crowd-sourcing. However, it struggles to achieve better recall and might miss several social bots. Finally, the experiments in this work focus primarily on the Sina Weibo platform. We expect to expand our analysis to other social media platforms such as 𝕏 (Twitter) or Reddit, in future work.

Ethics Statement

Research on misinformation and social bots is essential for countering online malicious content. This research demonstrates that social bots would amplify the spread of misinformation, enhancing echo chambers and manipulating public opinions. However, it may increase the risk of dual-use, where malicious actors may develop advanced social bots to spread misinformation. We will establish controlled access to ensure that the trained annotator checkpoints are only publicly available to researchers. Meanwhile, we will hide the privacy information in the dataset when we publish it.

Our models are trained on crowd-sourced data, which might contain social biases, stereotypes, and spurious correlations. Thus, our model would provide incorrect annotations. We argue that the predictions of our models should be interpreted as an initial screening, while content moderation decisions should be made with experts in the loop.

References

Zhila Aghajari. 2023. Adopting an ecological approach to misinformation: Understanding the broader impacts on online communities. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 417–420.

Joshua Ashkinaze, Eric Gilbert, and Ceren Budak. 2024. The dynamics of (not) unfollowing misinformation spreaders. In *Proceedings of the ACM on Web Conference 2024*, pages 1115–1125.

Dennis Assenmacher, Leon Fröhling, and Claudia Wagner. 2024. You are a bot!–studying the development of bot accusations on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 113–125.

Marina Azzimonti and Marcos Fernandes. 2023. Social media networks, fake news, and polarization. *European journal of political economy*, 76:102256.

Alexander Ratner Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3).

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. *arXiv preprint arXiv:2407.16670*.

Zijian Cai, Zhaoxuan Tan, Zhenyu Lei, Zifeng Zhu, Hongrui Wang, Qinghua Zheng, and Minnan Luo. 2024. Lmbot: distilling graph knowledge into language model for graph-less deployment in twitter bot detection. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 57–66.

Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. 2020. The role of bot squads in the political propaganda on twitter. *Communications Physics*, 3(1):81.

Peter Carragher, Evan M Williams, and Kathleen M Carley. 2024. Detection and discovery of misinformation sources using attributed webgraphs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 214–226.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Stefano Cresci. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Chaoqun Cui and Caiyan Jia. 2024. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 73–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yiqi Dong, Dongxiao He, Xiaobao Wang, Youzhu Jin, Meng Ge, Carl Yang, and Di Jin. 2024. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8354–8362.

Tuğrulcan Elmas, Rebekah Overdorf, and Karl Aberer. 2022. Characterizing retweet bots: The case of black market accounts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 171–182.

Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. 2022. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269.

Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021a. Satar: A self-supervised approach to twitter account representation learning and its application in bot detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3808–3817.

Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021b. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4485–4494.

Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. 2021c. Botrgcn: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 236–239.

Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What does the bot say? opportunities and risks of large language models in social media bot detection. *arXiv preprint arXiv:2402.00371*.

Shreya Ghosh, Prasenjit Mitra, and Preslav Nakov. 2024. Clock against chaos: Dynamic assessment and temporal intervention in reducing misinformation propagation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 462–473.

Haisong Gong, Weizhi Xu, Shu Wu, Qiang Liu, and Liang Wang. 2024. Heterogeneous graph reasoning for fact checking over texts and tables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 100–108.

Karish Grover, SM Angara, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Public wisdom matters! discourse-aware hyperbolic fourier co-attention for social text classification. *Advances in Neural Information Processing Systems*, 35:9417–9431.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Philipp Hartl and Udo Kruschwitz. 2022. Applying automatic text summarization for fake news detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2702–2713.

Chris Hays, Zachary Schutzman, Manish Raghavan, Erin Walk, and Philipp Zimmer. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection. In *Proceedings of the ACM web conference 2023*, pages 3660–3669.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H Andrew Schwartz, David H Epstein, Lorenzo Leggio, and Brenda Curtis. 2021. Bots and misinformation spread on social media: implications for covid-19. *Journal of medical Internet research*, 23(5):e26933.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2901–2912.

Zhen Huang, Zhilong Lv, Xiaoyun Han, Binyang Li, Menglong Lu, and Dongsheng Li. 2022. Social bot-aware graph neural network for early rumor detection. In *proceedings of the 29th international conference on computational linguistics*, pages 6680–6690.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Loukas Konstantinou and Evangelos Karapanos. 2023. Nudging for online misinformation: a design inquiry. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 69–75.

Zhenyu Lei, Herun Wan, Wenqian Zhang, Shangbin Feng, Zilong Chen, Jundong Li, Qinghua Zheng, and Minnan Luo. 2023. Bic: Twitter bot detection with text-graph interaction and semantic consistency. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Xiaojun Li, Xvhao Xiao, Jia Li, Changhua Hu, Junping Yao, and Shaochen Li. 2022. A cnn-based misleading video detection model. *Scientific Reports*, 12(1):6092.

Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024. Mcfend: a multi-source benchmark dataset for chinese fake news detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4018–4027.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Unsupervised keyphrase extraction by jointly modeling local and global context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 155–164.

Ying-Jia Lin, Chun-Yi Lin, Chia-Jen Yeh, Yi-Ting Li, Yun-Yu Hu, Chih-Hao Hsu, Mei-Feng Lee, and Hung-Yu Kao. 2024. Cfever: A chinese fact extraction and verification dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18626–18634.

Feng Liu, Zhenyu Li, Chunfang Yang, Daofu Gong, Haoyu Lu, and Fenlin Liu. 2024. Segcn: a subgraph encoding based graph convolutional network model for social bot detection. *Scientific Reports*, 14(1):4122.

Hui Liu, Wenya Wang, and Haoliang Li. 2023a. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796.

Yuhan Liu, Zhaoxuan Tan, Heng Wang, Shangbin Feng, Qinghua Zheng, and Minnan Luo. 2023b. Botmoe: Twitter bot detection with community-aware mixtures of modal-specific experts. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 485–495.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019.

Yen-Wen Lu, Chih-Yao Chen, and Cheng-Te Li. 2024. Dual graph networks with synthetic oversampling for imbalanced rumor detection on social media. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 750–753.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514.

Nikhil Mehta and Dan Goldwasser. 2024. An interactive framework for profiling news media sources. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 40–58.

Judith Möller, Robbert Nicolai Van De Velde, Lisa Merten, and Cornelius Puschmann. 2020. Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, 38(5):616–632.

Pedro Ramaciotti Morales, Manon Berriche, and Jean-Philippe Cointet. 2023. The geometry of misinformation: embedding twitter networks of users who spread fake news in geometrical opinion spaces. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 730–741.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *arXiv preprint arXiv:2402.16333*.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. *arXiv preprint arXiv:2405.16631*.

Lynnette Hui Xian Ng, Iain J Cruickshank, and Kathleen M Carley. 2022. Cross-platform information spread during the january 6th capitol riots. *Social Network Analysis and Mining*, 12(1):133.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865.

Zubair Qazi, William Shiao, and Evangelos E Papalexakis. 2024. Gpt-generated text detection: Benchmark dataset and tensor-based detection method. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 842–846.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2725–2732.

Lanyu Shang, Yang Zhang, Zhenrui Yue, YeonJung Choi, Huimin Zeng, and Dong Wang. 2024. A domain adaptive graph learning framework to early detection of emergent healthcare misinformation on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1408–1421.

Shuhao Shi, Kai Qiao, Jian Chen, Shuai Yang, Jie Yang, Baojie Song, Linyuan Wang, and Bin Yan. 2023. Mgtab: A multi-relational graph-based twitter account detection benchmark. *arXiv preprint arXiv:2301.01123*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Massimo Stella, Marco Cristoforetti, and Manlio De Domenico. 2019. Influence of augmented humans in online interactions during voting events. *PloS one*, 14(5):e0214210.

Hongyuan Su, Yu Zheng, Jingtao Ding, Depeng Jin, and Yong Li. 2024. Rumor mitigation in social media platforms with deep reinforcement learning. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 814–817.

Ling Sun, Yuan Rao, Yuqian Lan, Bingcan Xia, and Yangyang Li. 2023. Hg-sl: Jointly learning of global and local user spreading behavior for fake news early detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5248–5256.

Lingfeng Tan, Yunhong Wang, Junfu Wang, Liang Yang, Xunxun Chen, and Yuanfang Guo. 2023a. Deepfake video detection via facial action dependencies estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5276–5284.

Zhaoxuan Tan, Shangbin Feng, Melanie Sclar, Herun Wan, Minnan Luo, Yejin Choi, and Yulia Tsvetkov. 2023b. Botpercent: Estimating bot populations in twitter communities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14295–14312.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *IEEE Data Eng. Bull.*, 46(4):57–96.

Xiang Tao, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. Semantic evolvement enhanced graph autoencoder for rumor detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4150–4159.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. Gpt-who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024a. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Herun Wan, Minnan Luo, Zhixiong Su, Guang Dai, and Xiang Zhao. 2025a. On the risk of evidence pollution for malicious social text detection in the era of LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9731–9761, Vienna, Austria. Association for Computational Linguistics.

Herun Wan, Ningnan Wang, Xiang Zhao, Rui Li, Hui Yang, and Minnan Luo. 2024b. Fndpro: Evaluating the importance of propagations during fake news spread. In *International Conference on Database Systems for Advanced Applications*, pages 52–67. Springer.

Herun Wan, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, and Zhi Zeng. 2025b. Difar: Enhancing multimodal misinformation detection with diverse, factual, and relevant rationales. *arXiv preprint arXiv:2508.10444*.

Herun Wan, Jiaying Wu, Minnan Luo, Zhi Zeng, and Zhixiong Su. 2025c. Truth over tricks: Measuring and mitigating shortcut learning in misinformation detection. *arXiv preprint arXiv:2506.02350*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, pages 2452–2463.

Cheng-Jun Wang and Jonathan JH Zhu. 2019. Jumping onto the bandwagon of collective gatekeepers: Testing the bandwagon effect of information diffusion on social news website. *Telematics and Informatics*, 41:34–45.

Patrick Wang, Rafael Angarita, and Ilaria Renna. 2018. Is this the era of misinformation yet: combining social bots and fake news to deceive the masses. In *Companion Proceedings of the The Web Conference 2018*, pages 1557–1561.

Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. Msynfd: Multi-hop syntax aware fake news detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4128–4137.

Xiaofei Xu, Ke Deng, Michael Dann, and Xiuzhen Zhang. 2024. Harnessing network effect for fake news mitigation: Selecting debunkers via self-imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22447–22456.

Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. 2022. Botometer 101: Social bot practicum for computational social scientists. *Journal of computational social science*, 5(2):1511–1528.

Migyeong Yang, Chaewon Park, Jiwon Kang, Daeun Lee, Daejin Choi, and Jinyoung Han. 2024. Fighting against fake news on newly-emerging crisis: A case study of covid-19. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 718–721.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Zhiwei Yang. 2023a. Wsdms: Debunk fake news via weakly supervised detection of misinforming sentences with contextualized social wisdom. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1525–1538.

Yingguang Yang, Renyu Yang, Hao Peng, Yangyang Li, Tong Li, Yong Liao, and Pengyuan Zhou. 2023b. Fedack: Federated adversarial contrastive knowledge distillation for cross-lingual and cross-model social bot detection. In *Proceedings of the ACM Web Conference 2023*, pages 1314–1323.

Shu Yin, Peican Zhu, Lianwei Wu, Chao Gao, and Zhen Wang. 2024. Gamc: an unsupervised method for fake news detection using graph autoencoder with masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 347–355.

Xianghua Zeng, Hao Peng, and Angsheng Li. 2024. Adversarial socialbots modeling based on structural information principles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 392–400.

Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025. Imol: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024a. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19533–19541.

Fanrui Zhang, Jiawei Liu, Jingyi Xie, Qiang Zhang, Yongchao Xu, and Zheng-Jun Zha. 2024b. Escnet: Entity-enhanced and stance checking network for multi-modal fact-checking. In *Proceedings of the ACM on Web Conference 2024*, pages 2429–2440.

Huaiwen Zhang, Xinxin Liu, Qing Yang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. 2024c. T3rd: Test-time training for rumor detection on social media. In *Proceedings of the ACM on Web Conference 2024*, pages 2407–2416.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. 2024d. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16777–16785.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024e. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476.

Yuchen Zhang, Xiaoxiao Ma, Jia Wu, Jian Yang, and Hao Fan. 2024f. Heterogeneous subgraph transformer for fake news detection. In *Proceedings of the ACM on Web Conference 2024*, pages 1272–1282.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-stance: A large dataset for chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385.

Ke Zhou, Sanja Šćepanović, and Daniele Quercia. 2024. Characterizing fake news targeting corporations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1818–1832.

Ming Zhou, Wenzheng Feng, Yifan Zhu, Dan Zhang, Yuxiao Dong, and Jie Tang. 2023. Semi-supervised social bot detection with initial residual relation attention networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 207–224. Springer.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125.

A Details of MISBOT Dataset

A.1 Examples in MISBOT

Formally, an online information instance is represented as $A = \{s, I, V, \mathcal{G}_{repost}, \mathcal{G}_{comment}, \mathcal{U}, y\}$. The image set $I = \{I_i\}$ contains multiple images while the video set $V = \{V_i\}$ contains multiple videos. The repost graph $\mathcal{G}_{repost} = \{\mathcal{V}, \mathcal{E}, \mathcal{T}\}$ is a dynamic text-attributed graph (or tree) where the center node is the information content and another node $v \in \mathcal{V}$ denotes a repost text, $e = (v_i, v_j) \in \mathcal{E}$ denotes a repost relation connecting v_i and v_j, and $\mathcal{T} : \mathcal{V} \to \mathbb{R}$ denotes the published time of each node. $\mathcal{G}_{comment} = \{\mathcal{G}^i_{comment}\}$ denotes the comment graph set, where each comment graph $\mathcal{G}^i_{comment}$ is a dynamic text-attributed graph (or tree). Each comment graph is similar to the repost graph except for the center node, where the center node is a comment that directly comments on the information. Besides, a Weibo account instance is represented as $U = \{F, T, y\}$. The feature set contains *follower count*, *following count*, *status count*, *verified* (2 types), *svip* (2 types), *account type* (10 types), and *svip level* (6 types). The post set T contains the most recent five posts in the user timeline. We provide a piece of misinformation, real information, and a Weibo account example in Figure 10.

A.2 Management Center

The Weibo's official management center is a Weibo official. Here is the link: `https://service.account.weibo.com/?type=5&status=0`. If the users are logging into the platform for the first time, it will redirect to the Weibo homepage (`https://weibo.com/`). After logging in with a Weibo account, entering the platform again will lead to the right platform homepage. Figure 11 shows the overview of this platform, where we conceal private or unrelated information and translate the main language into English. If the users successfully log into this platform, they will view a similar website.

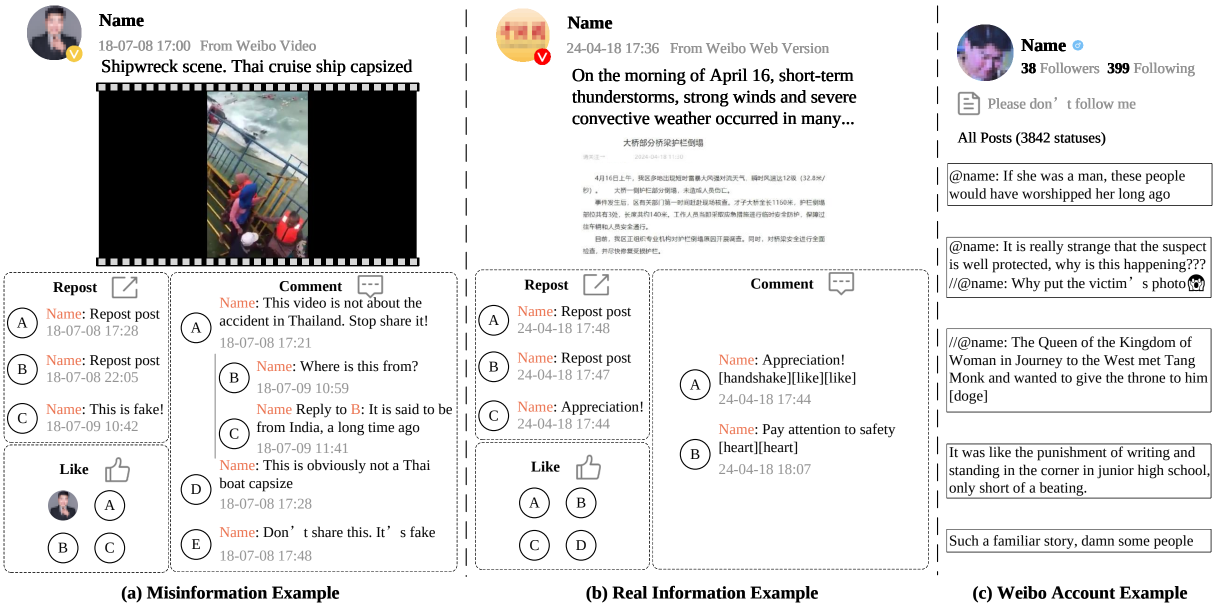


Figure 10: The examples in MisBot. We present (a) a misinformation example, (b) a real information example, and (c) a Weibo account example. We translate original information into English and conceal the private information.

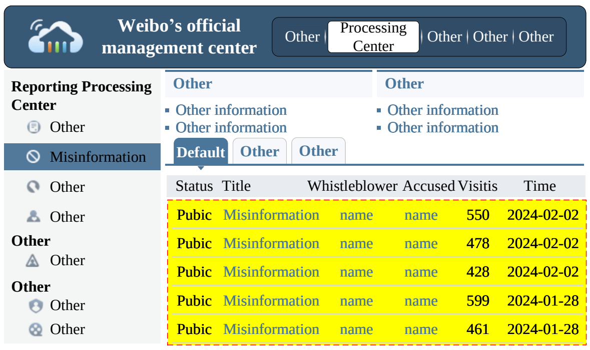


Figure 11: The overview of the Weibo's official management center. We conceal private or unrelated information and translate the main information into English. We highlight the misinformation items.

It is worth noting that the number of instances that each logged-in user can access per day is limited, so it took us about 10 months to collect all the misinformation on the platform.

A.3 Misinformation Example

After logging in to the platform, it mainly contains users' posts flagged as misinformation and a corresponding **judgment**. The platform moderators or police flag the misinformation and publish the **judgment**. We provide an example in Table 3. The judgment is the same for different pieces of misinformation on the same event.

Post flagged as misinformation: Recently, in xxx, a "naughty child" took scissors and cut off the hair of a female customer in a barber shop when no one was paying attention. After the female customer called the police and negotiated, the parents compensated 11,500 yuan.
Judgment: After investigation, it was found that the Weibo post claiming that "a woman's hair was cut off by a naughty child and her parents paid her 10,000 yuan in compensation" actually happened in May 2023, not recently. The respondent's speech is "outdated information" and constitutes "publishing false information"

Table 3: An example of misinformation and corresponding **judgment** (translated into English). The **judgment** provides a basis for identifying misinformation topics.

A.4 Entity Filter

We obtained 7,445 entities using the keyphrase extractor. We employ two strategies to filter out noisy entities:

- Frequency less than 10. These entities appear occasionally in misinformation and are unlikely to cause entity bias. We only focus on common entities that appear in large numbers in misinformation, so we need to ensure that they appear at a similar frequency in real information.

- The number of characters is 1. These entities

Homepage	Follower Count	Status Count	Discussion Count
https://weibo.com/u/1496814565	33.8 million	225.3 thousand	334.0 million
https://weibo.com/u/5044281310	32.6 million	163.2 thousand	573.0 million
https://weibo.com/u/1618051664	111.0 million	302.6 thousand	1.6 billion
https://weibo.com/u/1974808274	3.3 million	58.8 thousand	27.3 million
https://weibo.com/u/2028810631	107.0 million	166.4 thousand	469.0 million
https://weibo.com/u/2656274875	137.0 million	187.8 thousand	3.7 billion
https://weibo.com/u/1784473157	81.5 million	246.5 thousand	786.0 million
https://weibo.com/u/1642512402	62.4 million	224.4 thousand	410.0 million

Table 4: The information about the selected verified news accounts. We provide the homepage links of them. They have a huge number of followers and discussions.

might come from the noises of the keyphrase detector. Meanwhile, these entities may not contain enough semantic information.

After filtering, we obtained 1,961 entities. We believe these entities are common and contain rich semantic information. As a result, it would mitigate the effects of entity bias if real information also contains these entities.

A.5 Query Method

We mainly employ the official search function of the Weibo platform to search the given entity. Given an entity, the search function will return several posts containing the entity.

- **Verified news media**. After entering a specific account's homepage, we could use the search function to search posts in this account.

- **Trends on the platform**. Given an entity, such as *happy*, we collect posts in the trends using `https://s.weibo.com/weibo?q=happy&xsort=hot`.

Figure 12 presents an overview of these two search functions, where the red box illustrates the search function.

A.6 Verified Accounts

We employ 8 verified news accounts, and Table 4 presents the information about them. They have a red "verified" symbol. When an account has more than 10,000 followers and this account has been read more than 10 million times in the last 30 days, it can obtain the red "verified" symbol.

A.7 Source Credibility

Here we discuss the credibility of the two real information sources:

- **Verified news media**. These accounts are operated by legitimate news media and verified by the Weibo platform. Thus we believe this source is credible.

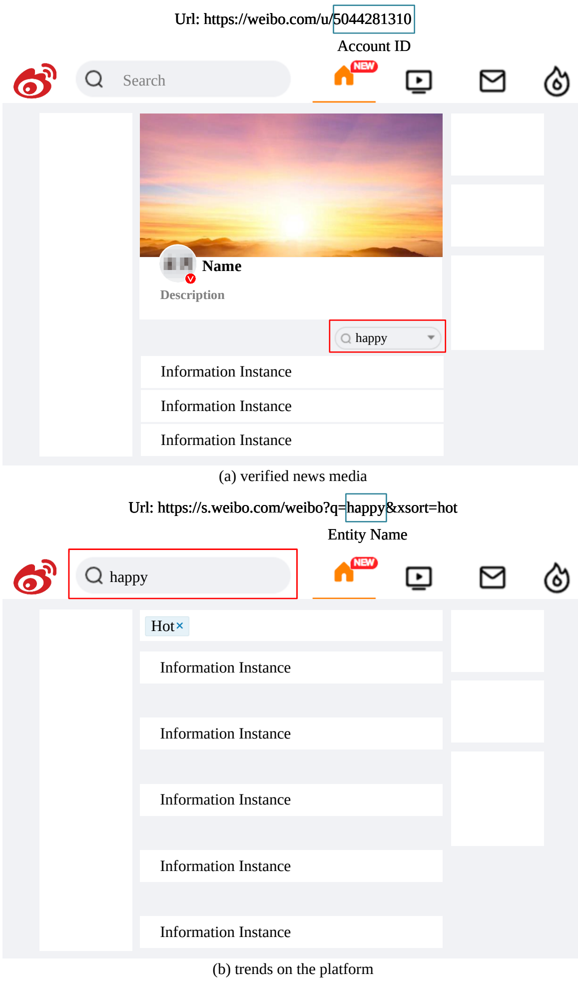


Figure 12: The overview of the two search pages. The red box presents the search functions.

- **Trends on the platform**. Weibo is a responsible social platform, where content moderators are efficient. As a result, the content moderation mechanism makes it easier to moderate posts with a lot of discussion. Because users would report the posts that they think are fake. After receiving reports, moderators only need to verify the post content instead of the whole discussion. It takes only a few days to moderate misinformation on the training. Meanwhile, the posts we collected are from one month ago in the trend. There is plenty of time for moderation.

A.8 Annotation Guideline

We first summarize the general criteria to identify a social bot on Weibo: (i) reposting or publishing numerous advertisements, (ii) devoted fans of a star publishing numerous related content, (iii) containing numerous reposting content without pertinence and originality, (iv) publishing numerous unverified and negative information, (v) containing numerous posts with the "automatically" flags, (vi) repeated posts with the same content, and (vii) containing content that violates relevant laws and regulations.

Based on the criteria, we write a guideline document for human annotators in Figures 13 and 14. Each human annotator must read this document before annotating.

A.9 Annotation Cost

Each human annotator is required to annotate 1,000 accounts plus 20 standard accounts. If a human annotator achieves more than 80% accuracy on the standard accounts, we will adopt the annotator's annotation. We will pay 200 yuan (about 28 dollars) for each qualified annotator. We recruited 315 annotators and, 300 are qualified. The crowd-sourcing takes about 60 days and costs 60,000 yuan.

A.10 Active Accounts

We focus on the active accounts in MISBOT and this paper. According to the human annotators' feedback and the characteristics of the Weibo platform. If an account publishes more than five posts with a length of no less than five characters in the timeline, then we consider this account active.

A.11 Expert Settings

In the training phase, we leverage three categories of social bot detectors as experts:

Feature-based Detectors We first preprocess the selected initial features to obtain the features for classifiers. For the **numerical** features (including *follower count*, *following count*, and *status count*), we employ z-score normalization:

$$z = \frac{x - \mu}{\sigma},$$

where x is the initial feature, z is the preprocessed feature, and μ and σ are the average and standard deviation in the training set. The average values are 5074.88, 420.59, and 1432.10, while the standard deviation values are 283145.61, 584.40, and 1373.91. For the **categorical** features (including *verified*, *svip*, *account type*, and *svip level*), we employ one-hot to obtain the initial representations. After that, we concatenate numerical and categorical representations to obtain the account representation $\boldsymbol{x}_f$. After that, We employ MLP layers, random forests, and Adaboost as detectors. We adopt three feature-based experts (three classic classifiers).

Content-based Detectors We employ *name*, *description*, and *posts* to identify social bots. We assuming the notation of *name* is $\boldsymbol{s}_{name}$, of *description* is $\boldsymbol{s}_{desc}$, and of *posts* is $\{\boldsymbol{s}^i_{post}\}_{i=1}^N$ (here are N posts). Given a text $\boldsymbol{s}$, we employ encoder-based language model to obtain the representation:

$$\boldsymbol{x} = \mathrm{LM}(\boldsymbol{s}).$$

For *posts*, we average the representation:

$$\boldsymbol{x}_{post} = \frac{1}{N}\boldsymbol{x}^i_{post}.$$

We feed the representations into an MLP layer to identify social bots. We employ the pre-trained parameters of the encoder-based language models and do not update the parameters. We employ the parameters in the Hugging Face for BERT[3] and DeBERTa[4]. We adopt six content-based experts (two encoder-based language models and three categories of texts).

Ensemble Detectors We first employ MLP layers to transfer the feature-based and content-based representations and concatenate them:

$$\boldsymbol{x} = \|_{i \in \{f, name, desc, post\}} \mathrm{MLP}(\boldsymbol{x}_i).$$

We adopt two ensemble experts (two encoder-based language models).

For all experts, we do not update the language model parameters. We set the *hidden dim* as 256, *learning rate* as 10^{-4}, *weight decay* as 10^{-5}, *batch size* as 64, *dropout* as 0.5, *optimizer* as Adam, *activation function* as LeakyReLU.

We do not employ graph-based detectors because neighbor information is hard to access on the Weibo platform and would cost a lot during the inference process. Besides, the automatic annotator already achieves acceptable annotation quality.

A.12 Temperature Settings

Temperature scaling is a post-precessing technique to make neural networks calibrated. It divides the logits (the output of the MLP layers and the input to the softmax function) by a learned scalar parameter,

$$p_i = \frac{e^{z_i/\tau}}{\sum_{j \in \mathcal{Y}} e^{z_j/\tau}},$$

where $\mathcal{Y}$ denotes the label set, p_i is the probability of belonging to category i. We learn the temperature parameter τ on the validation set. We conduct

[3]Here is the model link.
[4]Here is the model link.

Weibo Social Bot Annotation Guideline Document

Thank you for attending the Weibo social bot annotation.

This annotation aims to construct a large-scale Weibo social bot benchmark, where the main language is Chinese. The accounts are randomly selected from the Weibo platform, covering various account types.

You need to annotate 1,020 Weibo accounts. Given the homepage of a specific account, you need to determine whether it is a social bot or a genuine account.

Notes

If you are unsure about an account, remember the first impression is the most important.

There are 20 standard accounts that are easy to judge. As your accuracy on these accounts reaches 80%, your annotation will be accepted. If we accept your annotation, we will pay 200 yuan for you.

Guidelines

Here we provide a brief criteria and several examples:

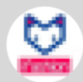
(a) Reposting or publishing numerous advertisements. Such accounts use Weibo to forward advertisements or product information in large quantities for commercial or profit purposes. If advertising-related posts are more than 40% of the total posts, they can be identified as social bots.

Pay attention to your skin
Pay attention to your skin
Pay attention to your skin

Containing numerous same ads.

[emoji][emoji]//phone ads
[emoji][emoji]//clothing ads
[emoji][emoji]//watch ads

Reposting numerous ads.



Name

Profile image, name, or description contains ads.

(b) Devoted fans of a star publishing numerous related content. Such accounts are mostly bought by stars to increase popularity and attract fans. They have obvious characteristics, where their homepage backgrounds are mostly photos or related information of a certain star, and more than 80% of their posts are related to the star.

Beautiful[emoji] @name
Gentle[emoji] @name
Sexy[emoji] @name

Mentioning the same star

Post about a star
[emoji]//Post about a star
Post about a star

Reposing or publishing posts about the same star

Figure 13: The overview of the guideline document, where we translate it into English. The human annotators are required to read this document before annotation.

a grid search from 0.5 to 1.5 with an interval of 0.001, obtaining the optimal value by minimizing the expected calibration error on the validation set.

B Details of Basic Analysis

B.1 Content Agreement Metrics

These three metrics are proposed to calculate the multi-modal content consistency, where a higher value means higher consistency. Formally, assuming each information instance is presented as (T_i, I_i) (here we only focus on the textual and image content). Meanwhile, the information set is represented as $\{(T_i, I_i)\}_{i=1}^{N}$ (misinformation set or real information set). Given an instance (T_i, I_i), we calculate **Text** as:

$$\textit{text}_i = \frac{1}{N}\sum_{j=1}^{N} \mathrm{cosine}(\mathrm{BERT}(T_i), \mathrm{BERT}(T_j)),$$

where $\mathrm{cosine}(\cdot)$ denote the cosine similarity function, $\mathrm{BERT}(\cdot)$ denote the BERT encoder[5]. We calculate **Similarity** as:

$$\textit{similarity}_i = \mathrm{cosine}(\mathrm{CLIP}(T_i), \mathrm{CLIP}(I_i)),$$

where $\mathrm{CLIP}(\cdot)$ denote the CLIP encoder[6]. We calculate **Image** as:

$$\textit{image}_i = \frac{1}{N}\sum_{j=1}^{N} \mathrm{cosine}(\mathrm{ViT}(I_i), \mathrm{ViT}(I_j)),$$

[5]Here is the model link.
[6]Here is the model link.

Weibo Social Bot Annotation Guideline Document (cont.)

(c) Containing numerous forwarding content without pertinence and originality. Such accounts simply repost others' posts.

(d) Publishing numerous unverified and negative information. Such accounts would publish shocking, negative, unconfirmed posts.

XXX was brutally murdered by the judge.

XXX was brutally murdered by the judge.

(e) Containing numerous posts with the “automatically” flags. Such accounts claim they are social bots in their name, description, or posts.

XXX**Bot**

self-proclaimed

Quick repost
Quick repost
Quick repost
From Weibo web version
Lots of automated behavior

(f) Repeated tweets with identical content. Such accounts would publish a lot of repetitive posts.

The same sentences
The same sentences
The same sentences
The same sentences

(g) Containing content that violates relevant laws and regulations. Such accounts would publish blood, violence, pornography content.

Figure 14: The overview of the guideline document (cont.).

where $\mathrm{ViT}(\cdot)$ denote the ViT encoder[7].

B.2 Image Distribution

To further explore the differences in image distribution between misinformation and real information, we check the categories of the image in information. We select four common categories: (i) *person*, (ii) *emoji pack*, (iii) *landscape*, and (iv) *screenshot*. We then investigate the sentiments of *person* and *emoji pack*, where *person* is realistic and *emoji pack* is virtual. The sentiments include neutral and non-neutral (angry, surprised, fearful, sad, and happy). To obtain the categories and sentiments, we employ pre-trained CLIP (Radford et al., 2021)[8] in zero-shot format. Figure 15 presents the image distribution of misinformation and real information. Images in real information tend to focus on people, while misinformation prefers to publish screenshots. Regarding sentiment, most of the images related to people in both real and misinformation are non-neutral, proving that information publishers tend to employ appealing pictures. For virtual images, emoji packs in real information are predominantly neutral, with a small partial being non-neutral. However, most emoji packs in misinformation are still neutral, significantly less than those in real posts. Furthermore, we analyze the correlation between the sentiment of images and text content (Appendix C.5), where 78.2% of real information contains images with the same sentiment as the text while only 34.1% of misinformation does. It further proves that misinformation and real information in MisBot are distinguishable.

B.3 Misinformation Detector

We propose a simple misinformation detector as Figure 16 illustrates. We employ multi-modal encoders to encode *content*, *repost*, *comment*, *image*,

[7]Here is the model link.
[8]Here is the model link.

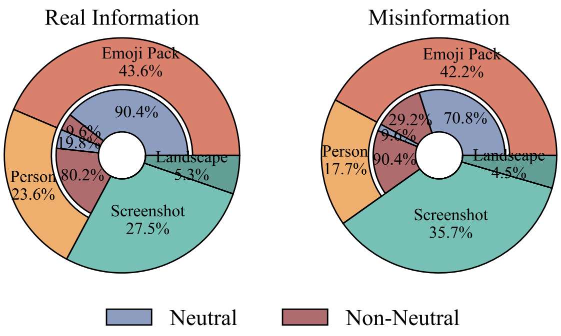


Figure 15: Image distribution of misinformation and real information, including categories and sentiments. Misinformation presents a different distribution from real information.

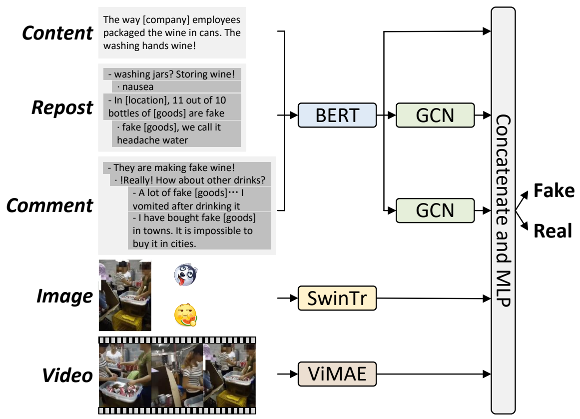


Figure 16: Overview of the misinformation detector, which employs multiple modality encoders to encode variance modalities and employs an MLP layer to identify misinformation.

and *video*. For *content*, we employ an encoder-based language model $\mathrm{LM}(\cdot)$[9] to encode content:

$$\boldsymbol{f}_{content} = \mathrm{LM}(s).$$

To encode *repost*, we employ the same language model $\mathrm{LM}(\cdot)$ to encode text-attributed node v_i and obtain $\boldsymbol{h}_{v_i}^{(0)}$. We employ L graph neural network layers to make each node interact:

$$\boldsymbol{h}_{v_i}^{(\ell)} = \underset{\forall v_j \in \mathcal{N}(v_i)}{\mathrm{Aggr}} (\{\mathrm{Prop}(\boldsymbol{h}_{v_i}^{(\ell-1)}; \boldsymbol{h}_{v_j}^{(\ell-1)})\}),$$

where $\mathcal{N}(v_i)$ denotes the set of neighbors of node v_i, $\mathrm{Aggr}(\cdot)$ and $\mathrm{Prop}(\cdot)$ are aggregation and propagation functions, where GCN (Kipf and Welling, 2017) is employed in practice. we finally employ the mean pooling operator as the $\mathrm{Readout}(\cdot)$ function to obtain the graph-level representation:

$$\boldsymbol{f}_{repost} = \mathrm{Readout}(\{\boldsymbol{h}_{v_i}^{(\ell)}\}_{v_i \in \mathcal{V}}).$$

To encode *comment*, we employ the same encoding method as *repost* to obtain the representation of each comment graph $\mathcal{G}_{comment}^{i}$ (Yang et al., 2023a). We then consider the average representations as the final representation:

$$\boldsymbol{f}_{comment} = \frac{1}{m} \boldsymbol{f}_{comment}^{i},$$

where m is the number of comment graphs. To encode *image*, we employ a pre-trained swin transformer[10] (Liu et al., 2022) $\mathrm{SwinTr}(\cdot)$ to obtain the representations of each image and adopt mean pooling to obtain the final representation:

$$\boldsymbol{f}_{image} = \mathrm{mean}(\mathrm{SwinTr}(I_i)),$$

where $\mathrm{mean}(\cdot)$ denotes the meaning operator. To encode *video*, we sample 256 frames from each video and resize each frame into 224×224. We employ pre-trained VideoMAE[11] (Tong et al., 2022) $\mathrm{VideoMAE}(\cdot)$. For each time step, we take 16 frames and set the interval to 12 frames. We could obtain:

$$\boldsymbol{f}_{video} = \mathrm{mean}(\mathrm{VideoMAE}(V_i)).$$

Finally, we concatenate them to obtain the representation of each user post:

$$\boldsymbol{f} = [\boldsymbol{f}_{content} \| \boldsymbol{f}_{repost} \| \boldsymbol{f}_{content} \| \boldsymbol{f}_{image} \| \boldsymbol{f}_{video}].$$

Given an information instance A and corresponding label y, we calculate the probability of y being the correct prediction as $p(y \mid A) \propto \exp(\mathrm{MLP}(\boldsymbol{f}))$, where $\mathrm{MLP}(\cdot)$ denote an MLP classifier. We optimize this model using the cross-entropy loss and predict the most plausible label as $\arg\max_y p(y \mid A)$. The hyperparameter settings of the baseline are presented in Table 5 to facilitate reproduction. We conduct ten-fold cross-validation to obtain a more robust conclusion. When split folds, we do not split misinformation from the same topic (Appendix C.1) into two folds to avoid data leakage.

B.4 Detector Ablation Study

We further design various variants of the misinformation detectors, removing certain components to explore which ones are essential for detection. We first remove each component except *content*. Then we design (i) w/o *Interaction* removing *comment* and *repost*; (ii) w/o *Vison* removing *image*

[9] Here is the model link.

[10] Here is the model link.

[11] Here is the model link.

Hyperparameter	Value	Hyperparameter	Value
BERT embedding dim	768	optimizer	Adam
GNN layers	2	learning rate	10^{-4}
GNN embedding dim	256	weight decay	10^{-5}
Video embedding dim	768	dropout	0.5
Image embedding dim	768	hidden dim	256

Table 5: Hyperparameter settings of the misinformation detector.

and *video*; and (iii) w/o *extra* only containing *content*. For each variant, we set the remove features as $\mathbf{0}$. For example, if we remove the *comment*, then we set $\boldsymbol{f}_{comment}$ as $\mathbf{0}$. We present the ablation study performance in Table 6. It illustrates that:

- The detector without *Extra* modalities suffers a significant performance decline, with an accuracy drop of 17.5%. It is more radical, often identifying information as misinformation and achieving high recall. It proves that extra modalities provide valuable signals to identify misinformation.
- The detector without *Interaction* drops to 77.3% on f1-score, illustrating the effectiveness of user reaction including comments and reposts. We speculate that user interactions could provide extra evidence and signals (Grover et al., 2022) to verify the information. Meanwhile, reposts provide more evidence than comments. We assume it is related to the algorithm of social platforms, where reposted messages could be spread more widely. Thus users tend to publish verified information when reposting.
- The detector w/o *Vision* only drops 2.2% on f1-score, where image and video information could not provide valuable evidence. Meanwhile, video information contributes the least, with the p-value of the t-test on accuracy being 0.015, which is not considered statistically significant. The text modalities dominate misinformation detection. We speculate that (i) annotators also consider text information when judging misinformation, introducing biases; and (ii) the pre-trained vision encoders struggle to capture signals related to identifying misinformation.

B.5 Expert Performance

We employ 11 social bot detectors as experts. Table 7 presents the performance and temperature of these experts. The performance of the automatic annotator is acceptable, proving the credibility of the annotations. Meanwhile, filtering in experts with

Models	Accuracy	F1-score	Precision	Recall
Vanilla	$95.2_{\pm 0.6}$	$92.3_{\pm 0.8}$	$93.7_{\pm 1.7}$	$91.0_{\pm 1.0}$
w/o *Comment*	$93.0^{\star}_{\pm 1.4}$ 2.3%↓	$89.5^{\star}_{\pm 1.6}$ 3.0%↓	$86.1^{\star}_{\pm 3.9}$ 8.1%↓	$93.3^{\dagger}_{\pm 1.6}$ 2.6%↑
w/o *Repost*	$89.4^{\star}_{\pm 2.0}$ 6.0%↓	$85.3^{\star}_{\pm 2.2}$ 7.6%↓	$76.8^{\star}_{\pm 4.0}$ 18.0%↓	$96.1^{\star}_{\pm 1.4}$ 5.6%↑
w/o *Image*	$94.3^{\star}_{\pm 0.5}$ 1.0%↓	$90.5^{\star}_{\pm 1.0}$ 1.9%↓	$95.1^{\dagger}_{\pm 1.2}$ 1.4%↑	$86.5^{\star}_{\pm 2.0}$ 4.9%↓
w/o *Video*	$95.0^{\dagger}_{\pm 0.7}$ 0.2%↓	$92.1^{\dagger}_{\pm 0.8}$ 0.3%↓	$93.0^{\dagger}_{\pm 1.9}$ 0.8%↓	$91.1^{\dagger}_{\pm 1.0}$ 0.2%↑
w/o *Interaction*	$81.6^{\star}_{\pm 4.5}$ 14.2%↓	$77.3^{\star}_{\pm 4.1}$ 16.2%↓	$64.4^{\star}_{\pm 5.9}$ 31.3%↓	$97.3^{\star}_{\pm 1.2}$ 7.0%↑
w/o *Vision*	$94.1^{\star}_{\pm 0.5}$ 1.1%↓	$90.3^{\star}_{\pm 1.0}$ 2.2%↓	$94.2^{\dagger}_{\pm 1.2}$ 0.4%↑	$86.8^{\star}_{\pm 2.1}$ 4.6%↓
w/o *Extra*	$78.5^{\star}_{\pm 5.2}$ 17.5%↓	$74.5^{\star}_{\pm 4.3}$ 19.3%↓	$60.6^{\star}_{\pm 6.0}$ 35.4%↓	$97.3^{\star}_{\pm 1.0}$ 6.9%↑

Table 6: Performance of the misinformation detector and variants. We report the mean and standard deviation of ten-fold cross-validation. We also report the performance changes and conduct the paired t-test with vanilla, where $\star$ denotes the p-value is less than 0.0005 and † denotes otherwise. This simple misinformation detector achieves ideal performance. Misinformation and real information are distinguishable with the help of user interactions.

an accuracy greater than 80% could improve the annotation precision. To obtain a higher precision, we set the likelihood threshold as 0.75, making sure that the annotator does not identify a genuine account as a social bot (with a precision of 97.6%).

C Details of Further Analysis

C.1 Cluster Algorithm

We cluster misinformation into different groups, where each group represents a topic or an event, based on the **judgment**. The main idea is that the judgments about the same event are very similar. Meanwhile, judgments about distinct events are very different. Formally, we assume the misinformation judgment set is $\{T_i\}_{i=1}^N$, where N is the number of misinformation judgments. Given a specific judgment T_i, we calculate the cosine similarities of BERT[12] representations:

$$s_{i,j} = \mathrm{cosine}(\mathrm{BERT}(T_i), \mathrm{BERT}(T_j)).$$

We sort the scores $\{s_{i,j}\}_{j=1}^N$ in descending order to obtain $\{s_{i,\tilde{j}}\}_{\tilde{j}=1}^N$. We then find the index $\hat{j}$ that maximize the gradient:

$$\hat{j} = \arg\max_{\tilde{j}} s_{i,\tilde{j}} - s_{i,\tilde{j}+1}.$$

[12]Here is the model link.

Experts	Accuracy	F1-score	Precision	Recall	Temperature
Feature-based Detectors					
MLP	73.5	49.6	90.9	34.1	1.125
Random Forest	71.7	58.0	67.0	51.1	–
Adaboost	69.5	59.8	60.2	59.5	–
Content-based Detectors (BERT)					
Name	74.5	54.3	86.4	39.6	1.468
Description	75.2	56.2	86.6	41.6	1.246
Posts⋆	80.4	72.4	78.4	67.2	1.286
Content-based Detectors (DeBERTa)					
Name	74.8	54.7	87.1	39.8	1.408
Description	75.2	58.7	80.9	46.1	0.972
Posts⋆	80.6	73.6	76.9	70.5	1.129
Ensemble Detectors					
BERT⋆	83.1	77.3	79.4	75.4	1.329
DeBERTa⋆	82.7	76.5	79.4	73.8	1.146
Annotator	85.0	79.5	83.3	76.1	–
All Expert	82.5	72.7	89.5	61.3	–
Annotator (0.75)	81.5	68.6	97.6	52.8	–

Table 7: The performance and temperature of the social bot detectors. The ⋆ indicates that we employ this expert in the final automatic annotator, and – indicates that temperature scaling is not suitable for this expert. The "All Expert" denotes the ensemble of all experts. The "Annotator (0.75)" denotes that we consider an account a social bot if the likelihood is greater than 0.75.

Keyword	Description
Fire Disaster	A place is on fire.
Dog Lost Notice	Someone offers a reward of 10 million yuan to find the dog.
Import	A country announced a ban on the import of another country's coal.
Typhoon	Does it feel like a disaster movie? A place is experiencing a typhoon.
Air Crash	The last two minutes of a place's air crash.
University Admission	A 19-year-old freshman girl in a city fell to her death and her roommate was recommended for undergraduate study.
BBQ	the woman beaten in the barbecue restaurant is dead.
Suicide	The woman who jumped from a place had her home disinfected and looted.
Child Trafficking	A 5-year-old son in a place was abducted near a bilingual kindergarten.
Domestic Violence	The man from a province is the stepfather, and I hope the relevant departments will save this poor child.

Table 8: The keywords and descriptions of 10 topics. We translate them into English and conceal the private information.

It means judgments with a similarity score greater than $s_{i,\hat{j}}$ are very similar to T_i and others are very distinct. Here we construct a relation from T_i to the judgments with a similarity score greater than $s_{i,\hat{j}}$. After that, we could obtain a directed graph. We consider each strongly connected graph as a misinformation graph.

C.2 Top-ten Topics

Table 8 presents the keywords and descriptions of the top 10 topics with the highest number of misinformation items. We employ BERT[13] to obtain the representations of misinformation.

C.3 Pairwise Scores

We conduct numerical analysis to prove that misinformation in the same cluster is similar, while misinformation in different clusters is distinct. We employ *semantic-level* and *token-level* pairwise scores. Formally, we assume there are N clusters (2,270 clusters), and the i-th cluster is represented as $\{T_k^i\}_{i=k}^{M_i}$, where M_i if the number of misinformation in this cluster. Given the i-th cluster and j-th cluster, the pairwise score s_{ij} is calculated as:

$$s_{ij} = \frac{1}{M_i M_j} \sum_{p=1}^{M_i} \sum_{q=1}^{M_j} \text{score}(T_p^i, T_q^j),$$

where $\text{score}(\cdot,\cdot)$ is the similarity function. For *semantic*, we employ the cosine similarity of BERT[14] representation. For *token*, we employ the jieba package[15] to tokenize Chinese sentences and calculate the ROUGE-L score. Since computing pairwise ROUGE-L is time-consuming, we randomly sample 10 pieces of misinformation in each cluster.

C.4 Score heatmap

Figure 17 presents the heatmap of the pairwise score, which illustrates that the values in the diagonal are much greater. It enhances our findings that: misinformation with the same topics has similar content and misinformation with different topics has distinct content.

C.5 Sentiments and Stances

To obtain the sentiments of social texts, we employ BERT trained on the EWECT dataset[16]. The sentiments include *neutral*, *happy*, *angry*, *surprised*, *sad*, and *fearful*. To obtain the stances of social texts, we employ BERT trained on the STANCE dataset (Zhao et al., 2023). The stances include *support*, *oppose*, and *neutral*.

[13] Here is the model link.

[14] Here is the model link.

[15] https://pypi.org/project/jieba/

[16] https://smp2020ewect.github.io/

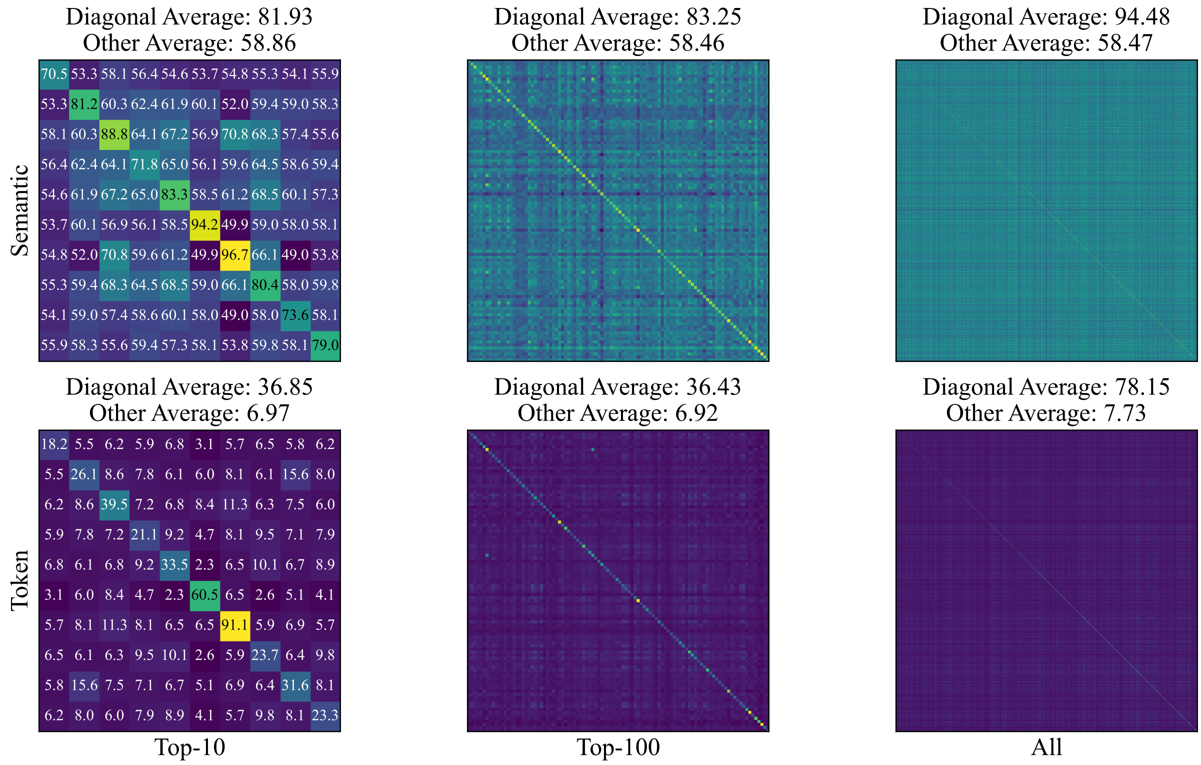


Figure 17: The pairwise score heatmap of *sematic* and *token* levels. The values on the diagonal are significantly larger than the rest. The “Top-10” means the 10 topics with the most misinformation instances, the “Top-100” means the 100 topics with the most misinformation instances, and the “All” means all misinformation instances.

C.6 Sentiment Distribution

Figure 18 illustrates the distribution of sentiments in different texts. An intuitive finding is that misinformation would publish more emotional content while real news would naturally report. However, whether in misinformation or real news, public reactions are always emotional. Comments in misinformation show more anger while real news shows more happiness, both of which are more emotional than reposts. We speculate that users are inclined to comment to express emotion.

C.7 Sentiment Variation

We introduce the variation measure to calculate the degree or extent to which public sentiment changes over the news spread. Given a specific information instance and its relation comment, we first calculate the function of the proportion of comments with neutral sentiment over time $f(x)$. We then determine the time series $[x_0, x_1, \ldots, x_n]$, where we set the interval as one hour. The variation is calculated as:

$$v_\Delta = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|.$$

C.8 Correlation Coefficient

To numerically explore the correlations between social bots and online public opinions, we calculate the following Pearson correlation coefficient:

- The number of social bots and the number of comments with non-neural stances: 0.6661.
- The number of social bots and the number of comments with non-neural sentiments: 0.6750.
- The ratio of social bots and the ratio of comments with non-neural stances: 0.2040.
- The ratio of social bots and the ratio of comments with non-neural sentiments: 0.2499.

The relatively high correlation coefficients indicate that social bots might influence public opinion.

C.9 Semantic Similarity

We explore the publishing behavior differences between social bots and genuine accounts. Here we explore whether accounts would publish similar content by introducing the semantic similarity score. Given an account with its posts in the timeline $\{T_i\}_{i=1}^{N}$, the semantic similarity is calculated

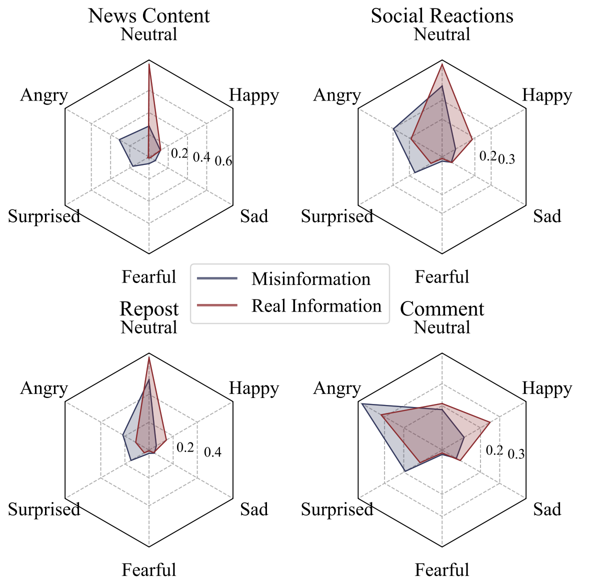


Figure 18: Sentiment distributions in different texts. Misinformation would publish emotional content while real information would publish more neutral content. Users would publish emotional content during the information spread.

as:

$$s = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\text{cosine}(\text{BERT}(T_i), \text{BERT}(T_j)).$$