Social Network Analysis and Mining (2022) 12:167
https://doi.org/10.1007/s13278-022-00992-8

ORIGINAL ARTICLE

# Do Twitter users change their behavior after exposure to misinformation? An in-depth analysis

Yichen Wang[1] · Richard Han[1] · Tamara Silbergleit Lehman[1,2] · Qin Lv[1] · Shivakant Mishra[1]

Received: 28 February 2022 / Revised: 14 October 2022 / Accepted: 17 October 2022 / Published online: 10 November 2022

**Abstract**
Social media platforms have been exploited to disseminate misinformation in recent years. The widespread online misinformation has been shown to affect users' beliefs and is connected to social impact such as polarization. In this work, we focus on misinformation's impact on specific user behavior and aim to understand whether general Twitter users changed their behavior after being exposed to misinformation. We compare the before- and after-exposure behaviors of Twitter users to determine whether they changed their tweeting frequency, tweets sentiment, usage of specific types of words, and the ratio of liberal/conservative media URLs they shared. Our results show that users overall exhibited statistically significant changes in behavior across some of these metrics. Through language distance analysis, we show that exposed users were already different from baseline users before the exposure. We also study the characteristics of several specific user groups, which include liberal/conservative leaning groups and multi-exposure groups. Furthermore, we study whether the users' behavior changes after exposure to misinformation tweets vary based on their follower count or the follower count of the tweet authors. Finally, we examine potential bots' behaviors and find they are similar to that of normal users.

**Keywords** Misinformation · Fake news · Twitter · User behavior

## 1 Introduction

Online social media has become increasingly popular in recent years and has been used to disseminate misinformation by users, sometimes intentionally, resulting in detrimental effects on our society. For example, some participants in the 2021 United States (US) Capitol riot said they were driven by online misinformation and conspiracy theories (Klepper 2021; Lemon 2021). As another example, misinformation is still driving people's vaccine hesitancy, especially during the COVID-19 pandemic (Bianco 2021). The spread of misinformation is a real threat to our society, as it can disrupt the public trust of legitimate news sources and undermine the political spectrum.

To combat misinformation, researchers have focused on two aspects: detecting misinformation and understanding its impact. To detect misinformation, researchers have built models making use of various information including content style, user profile and social context (Pérez-Rosas et al 2017; Shu et al 2018, 2019). To make it more amenable to the masses, academics have also proposed mechanisms to automate the fact-checking process (Hassan et al 2017; Ciampaglia et al 2015).

To understand misinformation's impact, researchers have investigated the spread pattern of misinformation (Del Vicario et al 2016; Vosoughi et al 2018), its negative effect on users' beliefs (Nyhan and Reifler 2010; Bessi et al 2015; Mocanu et al 2015), and its correlation with some social phenomena such as echo chambers and polarization (Ribeiro et al 2017). However, prior work has focused more on the misinformation's general social effect, and very little

✉ Yichen Wang
yichen.wang@colorado.edu

Richard Han
richard.han@colorado.edu

Tamara Silbergleit Lehman
tamara.lehman@colorado.edu

Qin Lv
qin.lv@colorado.edu

Shivakant Mishra
mishras@colorado.edu

[1] Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309, USA

[2] Department of Electrical, Computer and Energy Engineering, University of Colorado Boulder, Boulder, CO 80309, USA

work has been done to examine what and how specific user behavior is affected. We argue that it is crucial to study the details of specific behavioral changes after being exposed to misinformation. It can help us understand the process of how users succumb to misinformation and get affected negatively by exposure to misinformation. It can also help us identify specific user groups who are more likely to be vulnerable to misinformation and potentially even be radicalized. Some previous work studied the impact of COVID-19 related misinformation on users' vaccine intent (Loomba et al 2021), but they only focus on this specific type of misinformation. Limited work has focused on misinformation with broader topics and users' individual behavioral change.

In this paper, we have conducted a large-scale, quantitative analysis of Twitter user behavior after exposure to a known piece of misinformation. A user is considered to have been exposed to misinformation if they replies to a tweet carrying false information (*misinformation tweet*). We believe that the action of replying to a misinformation tweet is a much stronger indication of a user being exposed to misinformation and being influenced by it than other actions such as reading or liking a tweet. Specifically, to understand users' behavioral change, we seek to answer the following research questions (RQs):

- *RQ1*: Do the users who reply to misinformation tweets exhibit a change in their behavior after the exposure?
- *RQ2*: Are the behavioral changes different for users with different political leaning?
- *RQ3*: Does the change in behavior of users who reply to *multiple* misinformation tweets differ from other users?
- *RQ4*: Does the changes in user behavior of the users after being exposed to misinformation tweets vary based on the number of their followers or the number of followers of the tweet authors?
- *RQ5*: How do bots' replying behavior compare with normal users?

To identify misinformation tweets, we first obtained fact-checked misinformation excerpts from the well-known fact-checking website PolitiFact, and then queried Twitter to collect those tweets that contained these misinformation excerpts. Next, we collected the identities of all the users who replied to these misinformation tweets (named "target group" in the remaining part of this paper). To establish that any change in user behavior we observe is potentially due to exposure to misinformation tweets, we also built a user-controlled baseline group (named "baseline 1" in the remaining part of this paper) and an entity-matched baseline group (named "baseline 2" in the remaining part of this paper) for comparison. To identify whether there were significant changes in the tweeting behavior before and after exposure to misinformation, we selected objective behavioral metrics such as mean tweet frequency, mean sentiment score, and language usage distance among others. Then, we analyzed these behavioral metrics before and after exposure in both short-term (twenty-four hours before and after) and long-term (six months before and after). Furthermore, we compare users' behavioral change differences between liberal and conservative leaning users. Overall, this paper makes the following contributions:

- We introduce a dataset containing 372 misinformation tweets along with 21,071 users who replied to them and their tweets from six months before until six months after their reply.
- We reveal evidence of statistically significant changes in some behavioral features, such as the increase of frequency of tweets that users post, the ratio of liberal leaning media URLs they share, and the usage of emotion, swear and conflict related words. We do not observe such changes in the baseline user groups, indicating that there is a positive correlation between those behavioral changes and exposure to misinformation tweets.
- We do not find any significant change in user's overall tweet sentiment after being exposed to misinformation tweets. Using language distance analysis, we show that baseline users already had different language characteristics from the exposed users even before the exposure.
- We investigate users with a clear liberal or conservative leaning. We find that they have opposite changes in URL sharing behavior. We also find that liberal users have more change in the usage of emotion, swear and conflict related words.
- We investigate the group of users exposed to multiple misinformation tweets, i.e., they replied to more than one misinformation tweet. We find that these users' behavioral change is less than the users exposed to a single misinformation tweet, and that these users were already on a high activity level before the exposures.
- We find that exposed users with high and low follower counts exhibit similar behavioral change (and both the same as overall target group). Further, we find users exposed to misinformation tweets posted by users with high follower counts have more change in the usage of emotion, swear and conflict related words, compared with the users exposed to tweets posted by users with low follower counts.
- We find that bots that are exposed to the misinformation tweets are similar to normal users in terms of their replying behavior.

This paper is an extension of the preliminary version of the work by Wang et al (2021). The main limitation of that work is that it only considered tweet count and sentiment score to measure user behavior. In contrast, this paper significantly extends the preliminary work by (1) providing an in-depth description of the dataset and baseline building process, (2) incorporating more comprehensive behavioral features such as the usage of specific types of words and users' liberal/conservative media URLs sharing characteristics, (3) investigating bots' replying behavior, and (4) expanding the discussion about our work's implications and limitations. The usage of specific types of words can reflect users' social and psychological status, and the URLs sharing characteristics allow us to identify users' political leaning to better understand the impact that politics has on misinformation dissemination. The study on bots' behavior can help understand if bots play a role during the exposure to help with misinformation dissemination.

The intent of this paper is to identify and quantify correlation between exposure to Twitter misinformation and changed behavior in the exposed users where it exists, and not to establish causality, i.e., the paper does not claim that the changed behavior is caused by exposure to misinformation. Establishing causality would require further research. The Discussion section of the paper describes this in more detail.

The organization of the rest of our paper is as follows. In Sect. 2, we describe the background and related work on misinformation on social media. In Sect. 3, we present and explain the methodologies used to create the dataset and the user behavioral features under study. In Sect. 4, we present the results of the analysis of the research questions. Finally, in Sect. 5, we conclude the work and discuss its implications, limitations, and possible future directions.

## 2 Related work

Researchers have studied users and content on online social platforms extensively (Benevenuto et al 2009; Jin et al 2013), and it has been shown that online social media has become a major source of misinformation (Picchi 2018; Alba 2020; Javed et al 2020). A significant body of work has investigated the spread and detection of misinformation. Mustafaraj et al. described the spread process of fake news (Mustafaraj and Metaxas 2017). The diffusion process is also modeled by Tambuscio et al (2015). Making use of abundant data from a social network, Vosoughi et al. studied the spread pattern of fake news on Twitter from 2006–2017 and found that fake news spread farther, faster, deeper, and more broadly than true news (Vosoughi et al 2018). Vicario et al. studied the conspiracy news spreading on Facebook and found selective exposure is the primary driver of the diffusion (Del Vicario et al 2016). Social bots' role in misinformation spread was also studied (Shao et al 2018, 2017). To combat misinformation, scientists have studied and used a wide range of detection techniques. Journalists and investigators have built many manual fact-checking websites[1], and researchers have also explored automatic fact-checking methods (Ciampaglia et al 2015; Hassan et al 2017). Others have investigated automatic detection through content style (Pérez-Rosas et al 2017; Feng et al 2012; Zhou et al 2020), user profile (Shu et al 2018), and information propagation (Shu et al 2019; Ma et al 2018; Wu et al 2015). Our work differs from this body of research in that we focus on the consumer's behavioral changes after being exposed to misinformation instead of focusing on the producer side.

Another angle to study misinformation is to understand its impact on users and society. Psychologists and computer scientists have studied the impact of misinformation by looking at changes in user's beliefs and the overall social network. Researchers have shown that continued exposure to unsubstantiated rumors makes users more credulous (Bessi et al 2015; Mocanu et al 2015). Scientists have also shown that misbeliefs can persist after being exposed to misinformation (Nyhan and Reifler 2010). Loomba et al. focused on COVID vaccine related misinformation and found that users' vaccine intent decreased after exposure via qualitative analysis (Loomba et al 2021). Dutta et al. looked at the role of the political campaign during the 2016 United States Presidential Election by Russia's Internet Research Agency (IRA) among Twitter users (Dutta et al 2021). Researchers have investigated the correlation between misinformation and society polarization (Ribeiro et al 2017; Vicario et al 2019). Holme et al. also studied the influence on social network structure via simulations (Holme and Rocha 2019).

## 3 Methodology

In this section, we describe our data collection process. We explain how we established and collected the user groups we study, along with baseline user groups, and finally detail our decision to extend the behavioral features we used to analyze user's behavior.

### 3.1 Collecting tweets

The goal of this research is to understand behavioral changes of Twitter users after being exposed to misinformation tweets. Therefore, the first step is to identify the tweets that have misinformation content. As there is no "gold-standard" misinformation detection model, we resorted to the expert fact-checked news source as our "seed" to find the corresponding tweets. We choose PolitiFact as our misinformation source because PolitiFact has one section dedicated to fact-checking posts from social media platforms (Facebook) [2].

[1] https://www.snopes.com/; https://www.politifact.com/.

[2] https://www.politifact.com/personalities/facebook-posts/.

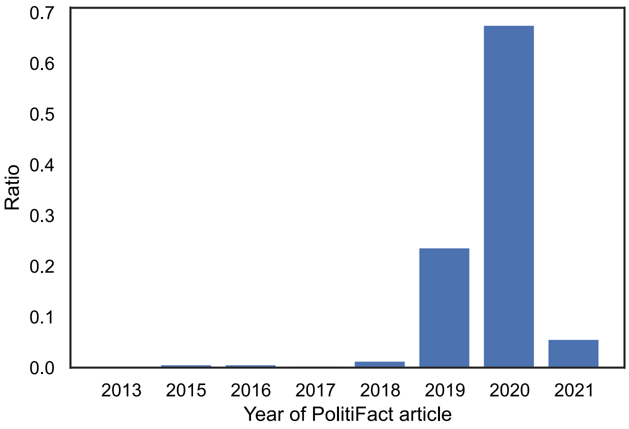
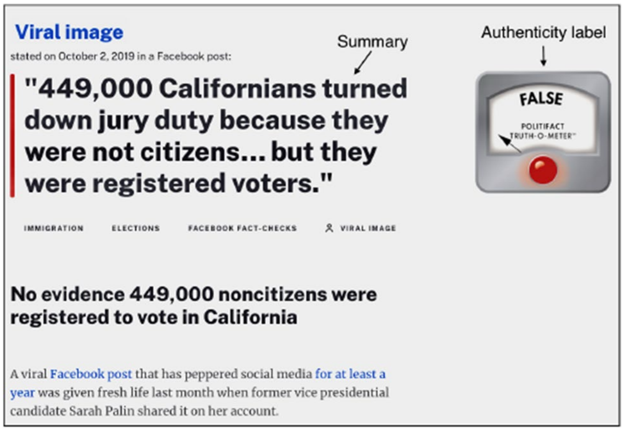


**Fig. 1** Ratio of PolitiFact articles broken down by the year they were posted



**Viral image**

stated on October 2, 2019 in a Facebook post:

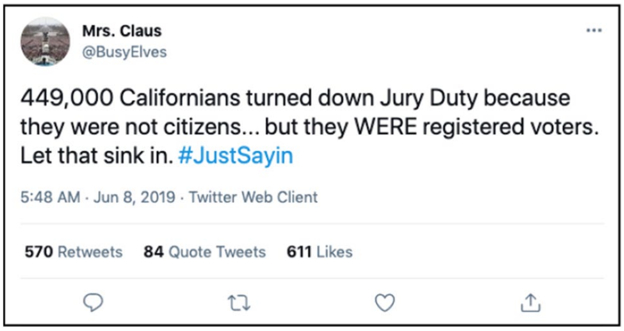
**"449,000 Californians turned down jury duty because they were not citizens... but they were registered voters."**

IMMIGRATION ELECTIONS FACEBOOK FACT-CHECKS VIRAL IMAGE

**No evidence 449,000 noncitizens were registered to vote in California**

A viral Facebook post that has peppered social media for at least a year was given fresh life last month when former vice presidential candidate Sarah Palin shared it on her account.

**Fig. 2** A Facebook post debunked by PolitiFact



**Mrs. Claus**
@BusyElves

449,000 Californians turned down Jury Duty because they were not citizens... but they WERE registered voters. Let that sink in. #JustSayin

5:48 AM · Jun 8, 2019 · Twitter Web Client

**570** Retweets **84** Quote Tweets **611** Likes

**Fig. 3** A tweet collected based on Fig. 2's article summary

We didn't directly study Facebook users because of their privacy policies. We need to collect each exposed user's activity data, but Facebook explicitly does not allow such collection to happen. Therefore, we collected Facebook misinformation and cross-searched them on Twitter, because misinformation from other popular social network platforms have a high probability of also showing up on Twitter. To achieve this, we crawled all the Facebook fact-checking articles from the PolitiFact website with authenticity level "pants on fire", "false", "mostly false" or "half true" dated between May 18, 2013 and Jan 31, 2021. Figure 1 shows the distribution of the fact-checking articles' release year, and most of them are in 2019 and 2020. Figure 2 shows an example of a fact-checked Facebook post on PolitiFact. For each fact-checking article, we used the provided summary as the search term to search for the corresponding misinformation tweets on Twitter up to 14 days after the fact-checking article was posted on PolitiFact. To avoid unrelated results, we disregarded the posts whose summary were less than seven words. We collected 1,119 debunked news articles from PolitiFact and misinformation in 442 of them could be found on Twitter. From the Twitter search results for each PolitiFact article, we extracted the top-5 tweets ranked by reply count. We then removed the tweets without reply and the ones that originated from fact-checking organizations, or that included any keyword regarding its veracity, e.g., "conspiracy theory", "debunk", and "fake news". After collecting all the tweets, we did a thorough verification to make sure our dataset only contained tweets with misinformed content. We manually removed all the tweets which were not misinformation and the users who replied to them, resulting in 399 tweets. Figure 3 shows an example of a tweet that we studied, which was retrieved using the summary in Fig. 2. In the next section, we will describe a further data cleaning process (bot removal) to build the final tweets dataset (372 tweets).

## 3.2 Collecting exposed users and baseline users

Figure 4 shows the overall flow of user data collection. We will elaborate the process in this section.

### 3.2.1 Collecting exposed users (Target Group)

For each of the collected tweets, we identified all the users who replied to it. Figure 5 shows the reply time with respect to the fact-checking article's posted time. Since the before and after analysis was performed on the users who replied to the tweets, and each user replied at a different time, the "zero" time for each user refers to the time of the user's first reply to the tweet. This method allows us to aggregate before and after behavior across different users who were exposed to misinformation tweets at different times. For each user, we collected all of their tweets starting at six months before their respective zero time and until six months after their respective zero time. A period of 30 days was used in place of a calendar month.

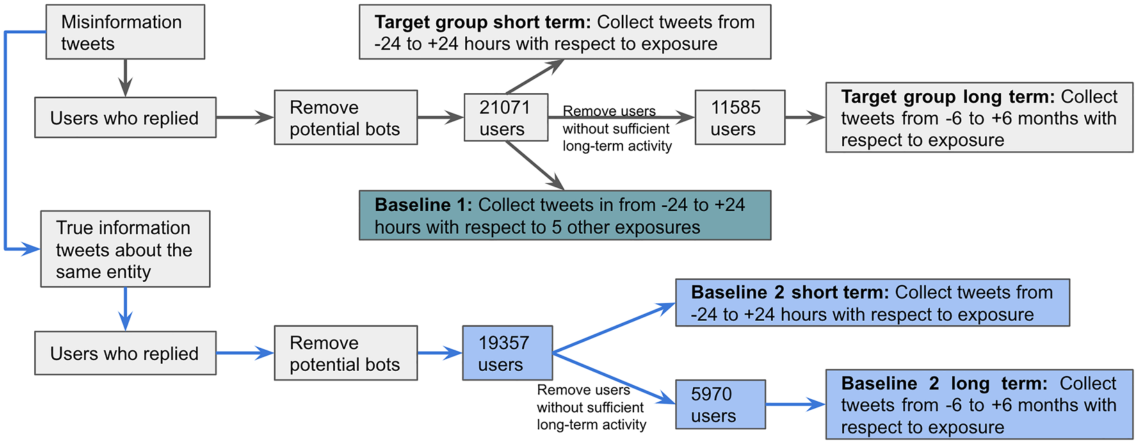

Misinformation tweets
Users who replied
Remove potential bots
21071 users
Target group short term: Collect tweets from -24 to +24 hours with respect to exposure
Remove users without sufficient long-term activity
11585 users
Target group long term: Collect tweets from -6 to +6 months with respect to exposure
Baseline 1: Collect tweets in from -24 to +24 hours with respect to 5 other exposures
True information tweets about the same entity
Users who replied
Remove potential bots
19357 users
Baseline 2 short term: Collect tweets from -24 to +24 hours with respect to exposure
Remove users without sufficient long-term activity
5970 users
Baseline 2 long term: Collect tweets from -6 to +6 months with respect to exposure

**Fig. 4** Overview of the data collection process

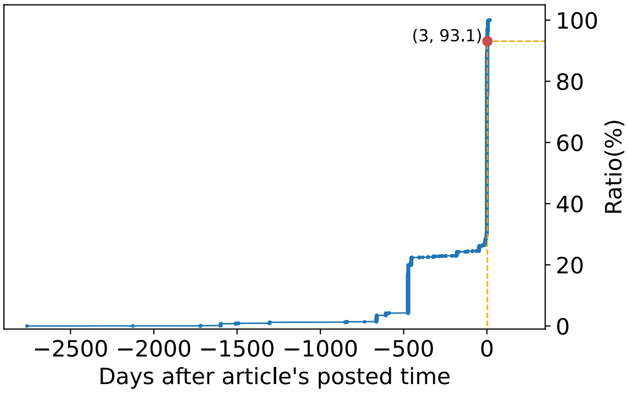


**Fig. 5** CDF plot of users' reply time with respect to PolitiFact fact-checking article's posted time. 93.1% of the users replied to the misinformation tweets no later than 3 days after the fact-checking articles were released on PolitiFact. There are also a few "old" misinformation which was already in Twitter years ago and got debunked in recent years, making the line starting from very long ago

In order to ensure the users under the purview of this study were legitimate users, we used Botometer[3] to remove users that were identified as potential bots. Botometer uses features from a user's profile along with machine learning to identify accounts primarily run with the help of automation software, and will return a score representing the probability that an account is run by a bot. Users with score above 0.5 were removed. There were 372 misinformation tweets with 21,071 replied users for target group after potential bot removal, which is the final dataset of the target group users we analyzed. The identified potential bots were studied in RQ5.

The long-term analysis includes only a subset of users who had activities throughout the whole 12-month period. The reason we had to exclude some users is that not all users had 6 months' of activity before or after the exposure. This way we only included the users for whom we had a complete 12 months of activity. This analysis also excludes users who joined Twitter within 6 months before the reply. There were 11,585 users for long-term analysis after the filtering.

### 3.2.2 Collecting baseline user groups

In order to ensure that the observed behavioral changes were sufficiently different from that of the general Twitter population and to exclude some unrelated factors from misinformation tweet content, we built two baseline groups for comparison. To ensure that the user behavioral change is potentially due to misinformation exposure, we tried to rule out other factors such as user's own personalities, and the topic (main content) of the (mis)information. As it is difficult to build a comparison group which has both user and content controlled, we built two baselines to control the two factors separately. The first baseline group is used to analyze the behavioral changes of the same users as the target group but being exposed to non-misinformation tweets, so that the users are controlled. The second baseline group is used to analyze a different group of users being exposed to true information but with similar topics as in the misinformation tweets, so that the tweet content is matched.

For the first baseline group, we used the same users as in the target group to understand the behavioral change when they are exposed to tweets without misinformation. To construct this baseline, we randomly collected 5 replies (exposures) to other tweets for each user and collected tweets before and after the exposure within the short-term (24 hours). The analysis on this baseline shows the average behavior of the 5 exposures. We selected 5 exposures because we are not able to confirm if other exposures are those of true news tweets, so we averaged multiple exposures

[3] https://botometer.osome.iu.edu/api.

to eliminate this possibility. This baseline is only used for the short-term analysis because there is a high possibility of overlapping periods for different exposures when looking at longer term behavior and it may interfere with the results.

For the second baseline group, we collected tweets from a different set of users who were exposed to content similar in subject matter to the target group but also true in nature. Because it is difficult to find related tweets without misinformation, we searched the tweets that only contained true news. To achieve this, we extracted the entity from each misinformation tweet's text content using Open Information Extraction (OpenIE) tool of Stanford CoreNLP [4]. Then, we collected all the recent tweets from known true news sources (Horne and Adali 2017) and excluded some questionable ones in recent years (e.g. The Guardian) and did the same entity extraction process. For each entity from misinformation tweets, we chose a tweet with the same entity from the true-information tweet group with similar reply count. Due to the limitation of our crawling tools, only 3,200 most recent tweets could be fetched for each source, thus not all misinformation tweets could be matched. For the remaining unmatched misinformation tweets, we used their entities as the search term to search for related tweets. To ensure the tweet's veracity, we only considered tweets posted by verified accounts and gave priority to the known true news sources. We then selected tweets whose reply count was close to entity-matched misinformation tweets. Finally, we performed the same user scraping process as before to get users' tweets from 6 months before until 6 months after the exposure, and then removed potential bots. Figure 4 shows the overall process of user data collections and the actual number of users we considered for the analysis. There are far fewer users in baseline 2 group for the long-term analysis because many of the tweets collected from reliable sources were very recent ones and the exposed users did not have 6-months worth of activities after exposure.

We excluded retweets and favorite tweets in this work since the tool we used to scrape Twitter has a time constraint for these features. We used Twint [5], a Twitter-scraping Python library, to search and collect the tweets as described above. Twint can only retrieve the most recent retweets and favorite tweets. When working with non-recent data, it is unlikely that a majority of favorite tweets from said period will be reachable. Twint is also limited in the state of the accounts it is able to retrieve. It is unable to retrieve deleted account data, and tweets posted when the corresponding account is deleted or private. To ensure completeness and fairness of our dataset, we decided to exclude retweets and favorite tweets as well as accounts for which tweets could not be reached.

### 3.3 Features

We analyzed the before and after user behavior through several specific metrics: tweet frequency, sentiment score, subjectivity, language usage distance. All features were studied hourly (short-term) and/or monthly (long-term). They were averaged to form the hourly or monthly statistics for each user.

*Tweet Frequency* was determined for each user by counting the total number of tweets posted by the user within the bounds of a 1-hour or 1-month period. This count includes tweets posted by the user and replies to other accounts during that time period.

*Sentiment Score* was calculated using VADER, a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiment expressed in social media (Hutto and Gilbert 2014). A sentiment score within the range [-1, 1] was assigned to each tweet based on its content. A score close to -1 indicates a highly negative sentiment while a score close to 1 indicates a highly positive sentiment.

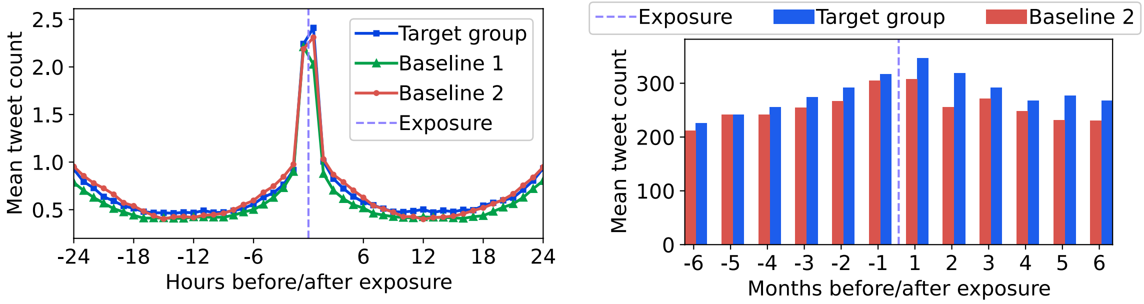
*LIWC word categories* (Boyd et al 2022) were used to measure the change of psychological state of users before and after exposure. Linguistic Inquiry and Word Count (LIWC) is a gold-standard lexicon containing various sets of word categories that were created to capture people's social and psychological states. We calculated the percentage of words belonging to several word categories in each user's language. Specifically, we focused on cognitive thinking related words which can reflect users' critical thinking, emotion related words, swear words, and words reflecting conflict with other people. As there is not enough usage of these words in users' language in the 24-hour period, we only analyzed each user's average monthly usage percentage of these words. Therefore, only the target group and baseline 2 group were compared. We removed all the mentions ("@user") and URLs from the tweets in the preprocessing step.

*Language Usage Distance* was used to evaluate the difference of language between two groups of tweets text. Similar to prior work (Hessel et al 2016; Althoff et al 2016), we adopted the Jensen-Shannon Divergence (Fuglede and Topsoe 2004) to measure the unigram difference (hourly and monthly) in tweets as the language distance. A larger distance indicates a larger difference in language (word) usage. We removed all the mentions, URLs and stopwords from the tweets, and stemmed the words as the pre-processing steps.

We used dependent sample t-test to assess the statistical significance of the results for each feature. We used this type of test because the before and after samples are not independent of each other. We aggregated the users' hourly/

[4] https://stanfordnlp.github.io/CoreNLP/.

[5] https://github.com/twintproject/twint.



(a) Hourly mean tweet count in ±24 hours.

(b) Monthly mean tweet count in ±6 months.

**Fig. 6** Mean tweet count comparison

monthly feature before and after the exposure to conduct the tests.

### 3.3.1 Inferring political leaning

To infer a user's political leaning, we applied a similar approach used in Luceri et al (2019), which relies on the political leaning of the URLs of media outlets they shared. We referred to a list of liberal and conservative media outlets' URLs compiled by Media Bias/Fact Check [6]. Liberal and liberal-center bias outlets are combined (937 URLs), as well as conservative and conservative-center bias outlets (592 URLs). We then extracted URLs from users' tweets and matched them with the media URLs. As most URLs on Twitter get shortened (e.g., *https://t.co/G84PK2l1Wl*), we need to expand it, therefore we only considered the URLs with top 1 million frequency in the tweets.

To label a user as liberal or conservative, we used a rule that is based on the ratio of liberal/conservative URLs they shared before the exposure. For each user, we identified all the liberal, conservative, and neutral media URLs they shared, and calculated the ratio for each type of URL. If a user shared more liberal URLs than conservative and neutral URLs, and the ratio of liberal URLs is at least 10% more than that of the conservative URLs, this user is labelled as liberal, and vice versa. If the difference is less than 10%, the user is labelled neutral.

The monthly ratios of liberal, conservative and neutral URLs were also used as behavioral features in our analysis. We did not analyze the ratios in shorter term because these identified URLs only cover a small number of users' shared URLs and we do not have enough URL sharing data for the 24-hour granularity.

## 4 Results

In this section, we present the results of our analyses broken down by each of the research questions. The main findings include: (1) the target group significantly increased their tweeting frequency and changed word usage in several word categories; (2) liberal leaning users show differences compared with conservative leaning users; and (3) users exposed to multiple misinformation tweets show different and larger behavioral changes; (4) user language has more change after exposure to misinformation authored by popular users; and (5) bots are similar to normal users in terms of their replying behavior.

### 4.1 RQ1: Do users who reply to misinformation tweets exhibit a behavioral change after exposure?

*The target group significantly increased tweeting frequency and decreased their sharing of liberal leaning URLs following their exposure to the misinformation tweets compared with the baseline users.* As shown in Fig. 6a in the short-term analysis all three groups' tweeting frequency has a 24-hour periodic change. The target group's tweeting frequency increases significantly (0.66 vs. 0.68***[7]) during the 24-hour period after the exposure, while baseline 1 group's decreases significantly (0.60 vs. 0.59***). For baseline 2 users, there is no significant tweeting frequency change (0.67 vs. 0.68, $P = 0.24$). Note that we also analyzed the behavior for a 72-hour period and it shows similar patterns. Due to space limitations, we only report the 24-hour analysis.

The long-term tweet frequency has a similar pattern, as shown in Fig. 6b. Although monthly tweet count increases

[6] https://mediabiasfactcheck.com/.

[7] *Note:* Throughout this paper, the comparison of the feature value before and after exposure is shown as $\text{Fea}_{\text{before}}$ vs. $\text{Fea}_{\text{after}}$, and $P$-values are indicated by the stars: ***: $P < 0.001$, **: $P < 0.01$, *: $P < 0.05$.

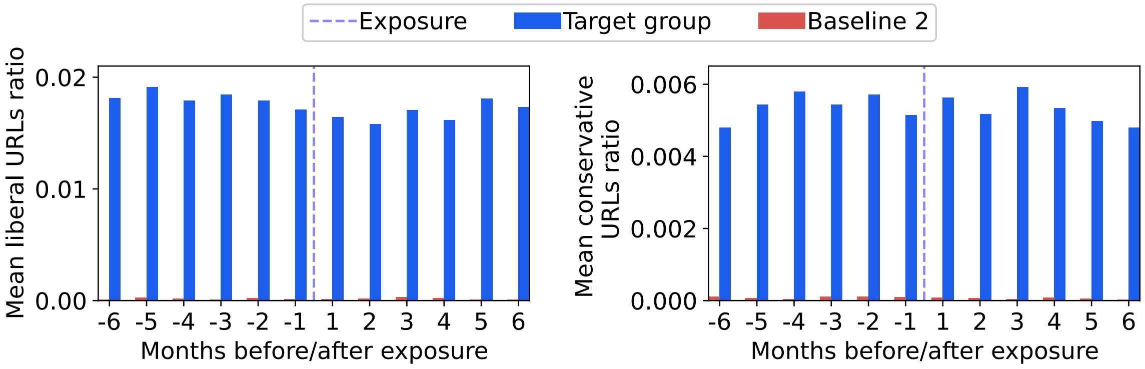
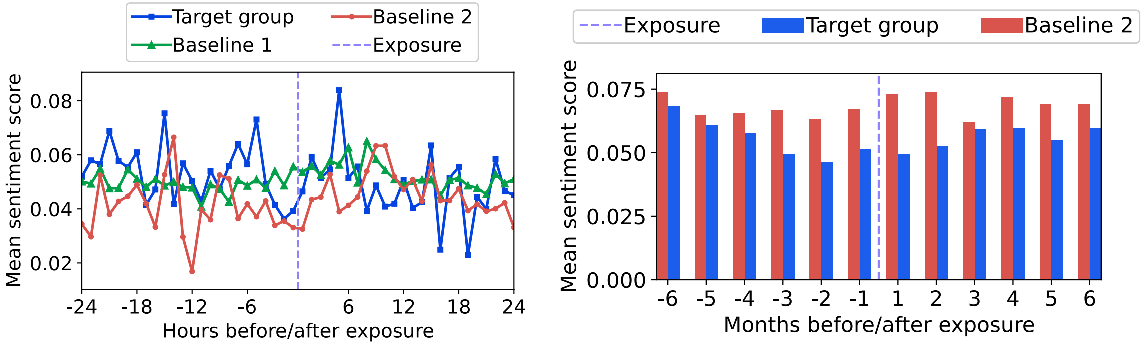


**Fig. 7** Liberal and conservative media URLs sharing ratio comparison



(a) Hourly mean sentiment score in ±24 hours.

(b) Monthly mean sentiment score in ±6 months.

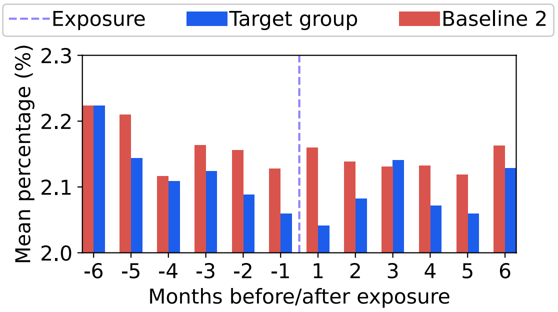
**Fig. 8** Mean sentiment score comparison

for both groups, the target group's tweet count increases more significantly. The baseline 2 group's change is significant (253.5 vs. 257.6*), but the change amount and significance level are much weaker than that of the target group (267.7 vs. 294.9***).

Target group users' sharing ratio of liberal media URLs also has a change, where the monthly sharing ratio decreases (0.018 vs. 0.017***). As a comparison, the baseline 2 group does not have a significant change (0.00016 vs. 0.00017, $P = 0.80$). This change shows a potential shift of users' preference towards media in terms of their political leaning, which partially aligns with recent work on the association between fake news exposure and decline in mainstream media trust (Ognyanova et al 2020). As shown in Fig. 7, target group users shared way more liberal/conservative media URLs than baseline 2 group in the 12-month period, which potentially suggests that these two groups of users are different people in nature.

*The target group users did not change their sentiment significantly, but changed language usage in several word categories.* To measure users' language change, we first examined their tweet sentiment score. Sentiment score of all the 3 groups in the short-term does not change significantly (Fig. 8a). As shown in Fig. 8b, the target group's sentiment score does not change significantly in long-term either (0.056 vs. 0.056, $P = 0.85$), while baseline 2 group's sentiment score has a small increase (0.067 vs. 0.070***). We argue that this is because a user's sentiment does not necessarily change in one direction (only increase or decrease) after the exposure. A person may express the same stance/opinion toward a tweet by using either negative or positive language (Mohammad et al 2017; Aldayel and Magdy 2019), which would not change the average sentiment score significantly.

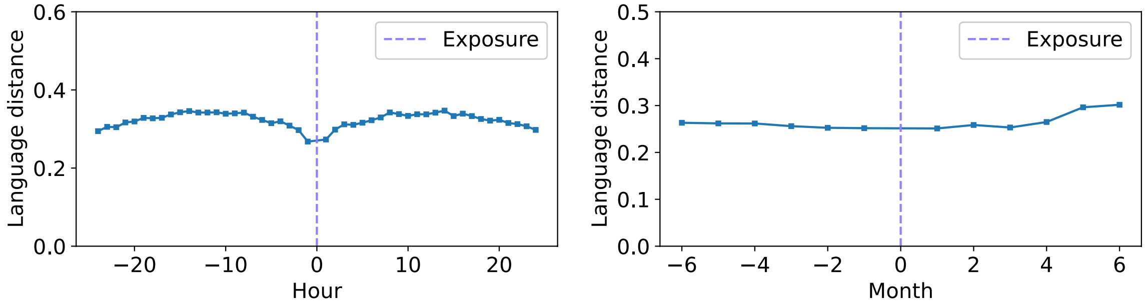
The usage of words (in percentage) of different categories was examined to analyze their linguistic and psychological status. Since only monthly usage data could be analyzed, only the target group and baseline 2 group were compared. We first found that target group users use less emotional words (2.12 vs. 2.09%***), while baseline 2 users do not have this change (2.17 vs. 2.14%, $P = 0.10$). Next, we found that target group users use more swear words (0.62 vs. 0.65%***) and words reflecting conflict with others (0.52 vs. 0.54%***), while baseline 2 users do not have these



(a) Emotion related words usage.

(b) Swear words usage.

**Fig. 9** Mean monthly usage ratio change of emotion and swear words



(a) Language distance in ±24 hours.

(b) Language distance in ±6 months.

**Fig. 10** Language distance between target group and baseline 2

changes (0.86 vs. 0.86%, $P = 0.93$ for swear words, 0.39 vs. 0.37%*** for conflict words). This demonstrates that although the target group users do not change their language sentiment, they are more reluctant to express emotion, i.e., both positive and negative emotions in their language decreased. Meantime, their language becomes more aggressive, reflected from swear words and words related to conflict with others. In addition, we found that both target group and baseline 2 users decrease the usage of words reflecting cognitive thinking (10.76 vs. 10.71%*** for target group, 9.16 vs. 9.10%*** for baseline 2). Figure 9 shows the change of emotional and swear words.

To understand these results further, we compared the target group with the baseline 2 group by their language distance. We calculated the language distance for each hour/month before and after the exposure between the target and the baseline 2 group. As shown in Fig. 10, the language distance of the target group with the baseline group is stable before and after the exposure, with a slight increase starting from the fifth month after exposure. This observation indicates that the target and the baseline 2 group users already have different language characteristics even before their respective exposures and that this difference does not change after the exposure. The large differences of URL sharing shown in Fig. 7 also supports this claim. This indicates that the misinformation and true information tweets attract users with different characteristics.

### 4.2 RQ2: Are the behavioral changes different for users with different political leaning?

We extracted the URLs from the 6-month period before the exposure and conducted the political leaning inference. 2,631 liberal users and 582 conservative users were identified from the target group using the inference method described in Sect. 3.3.1.

*Liberal users are different from conservative users in some behaviors, especially in their sharing of liberal/conservative media URLs, and specific word usage.* The first difference lies in their long-term tweet frequency. Liberal users increase their tweet frequency in the long-term (436.8 vs. 473.6***), which is the same trend as the overall target group, but conservative users' tweet frequency had no significant change in long term (394.2 vs. 405.5, $P = 0.12$). Their tweets' sentiment score has the same trend as that of the overall target group.

The users' URL sharing behavior has a large difference across political leaning. For liberal media URL sharing, liberal users have a significant decrease (0.062 vs. 0.048***), while conservative users are the opposite (0.010 vs.

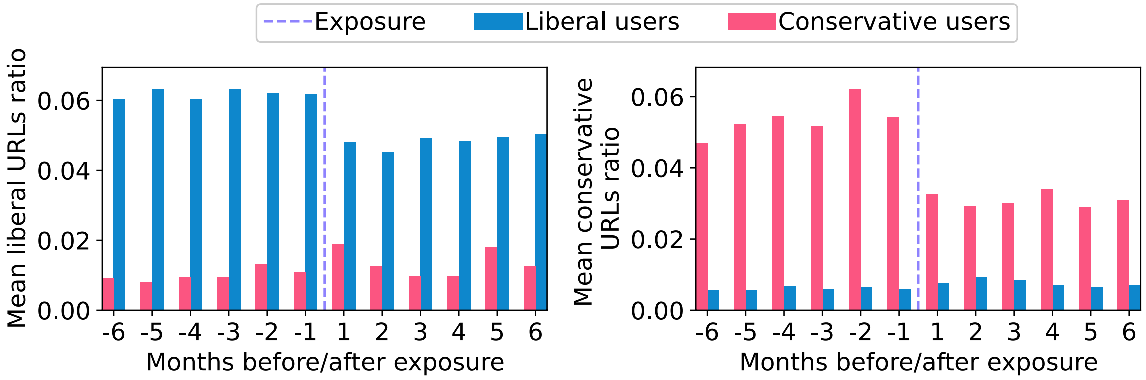
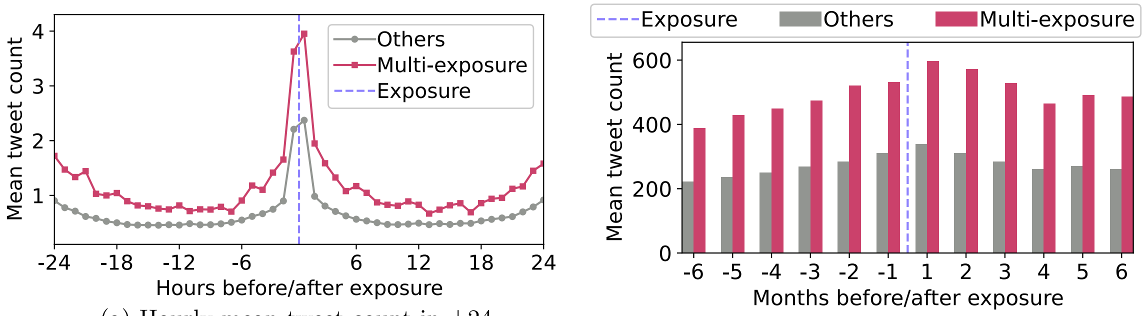


**Fig. 11** Liberal and conservative media URLs sharing ratio comparison of liberal and conservative users



(a) Hourly mean tweet count in ±24 hours.

(b) Monthly mean tweet count in ±6 months.

**Fig. 12** Mean tweet count comparison with respect to exposure count

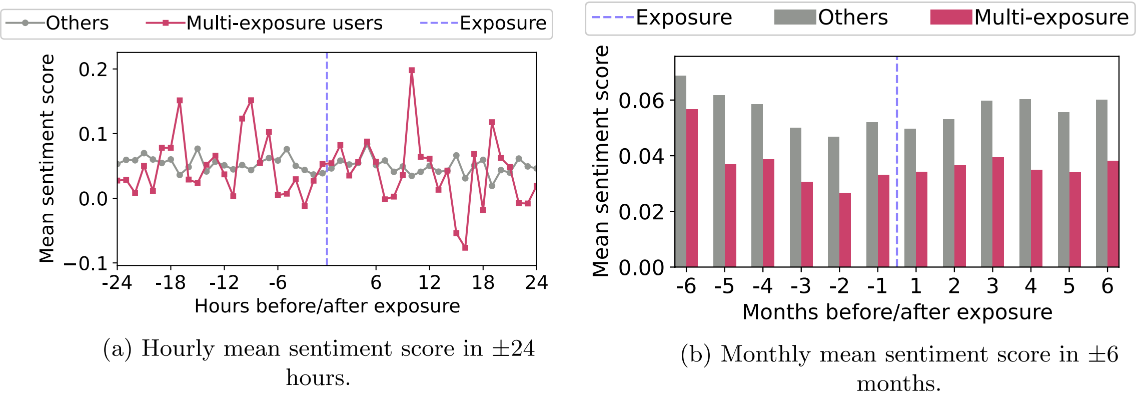
0.014**). For conservative media URL sharing, the results show the reverse case, where liberal users have a significant increase (0.006 vs. 0.008***), while conservative users have a significant decrease (0.054 vs. 0.031***). Figure 11 shows the results. We also compared these users with the liberal and conservative leaning users in baseline 2, where the baseline 2 users' sharing of those URLs didn't have significant changes, and they generally post much fewer URLs of liberal/conservative media.

They also differ in the word usage across several categories. Although both liberal and conservative users decrease their cognitive thinking related words usage significantly, liberal users' word usage in other categories change more differently than that of the conservative users. Liberal users decrease emotion related word usage (1.85 vs. 1.82%*) significantly, while conservative users do not. Liberal users increase swear words (0.48 vs. 0.50%***) and conflict related words (0.61 vs. 0.63%**) significantly, while conservative users do not. This result suggests that liberal users' emotion and aggression in language fluctuate more easily than that of conservative users after exposure to misinformation.

### 4.3 RQ3: Does the change in behavior of users who reply to *multiple* misinformation tweets differ from other users?

*Multi-exposure users show a significant change of tweet count in long-term but not in short-term, and behave more "unstable" than other users.* We consider multi-exposure users to be the ones who replied to at least two misinformation tweets. Although users may reply to other misinformation tweets, in this work we only consider the exposure to our collected tweets. There are 504 users in this group.

As shown in Fig. 12a, the tweet count for multi-exposure users does not have significant change in the short-term (1.14 vs. 1.17, $P = 0.37$), while other users' (single-exposure) increase is statistically significant (0.64 vs. 0.67***). In the long-term, multi-exposure users still have an increase in tweet count (465.0 vs. 523.0***), which has the same trend as that of the single-exposure users (261.2 vs. 287.3***). In terms of their language features, we do not observe significant sentiment change for both the multi-exposure user group and others, which is the same as the overall target group. Multi-exposure users' word usage on cognitive thinking, emotional, swear and



(a) Hourly mean sentiment score in ±24 hours.

(b) Monthly mean sentiment score in ±6 months.

**Fig. 13** Mean sentiment score comparison with respect to exposure count

conflict related words also have the same change trend as others, and is the same as the overall target group.

From the comparison between the multi-exposure and single-exposure groups, it is shown that the multi-exposure group generally posts more tweets (Fig. 12) with more volatile sentiment (Fig. 13). This difference is stable across the 12 months (their long-term sentiment score is lower because averaging the volatile score in a longer term results in average monthly score closer to 0). We conclude that the multi-exposure users are already in a "high-level mood" and their short-term change is less than that of the single-exposure users, who are rising from a relatively lower level, and generally they are "unstable" compare with the single-exposure users in the language usage.

### 4.4 RQ4: Does the change in the behavior of the users after being exposed to misinformation tweets vary based on the number of their followers or the number of followers of the tweet authors?

We conducted two analyses for this research question, where the first is to understand if the exposed users behaved differently when their follower count is different, and the second is to understand if the exposed users behaved differently when the misinformation tweet authors' follower count is different. We separated the exposed users into low-follower count and high-follower count groups, where 240 was chosen to be the threshold for high and low follower count because 240 divided the users fairly well into two halves. Using the same idea for the misinformed tweets authors, 5400 was chosen as it separates the authors into two halves. Figure 14 shows the distribution of the followers. As a result, there were 13,797 and 9,325 users exposed to tweets authored by high-follower count users for short and long-term respectively, while there were 1,597 and 1,004 users exposed to tweets authored by users with low-follower count for short and long-term respectively.

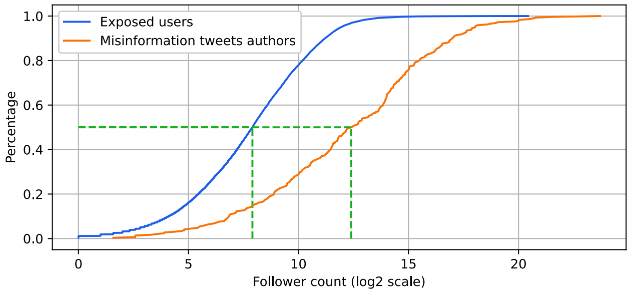
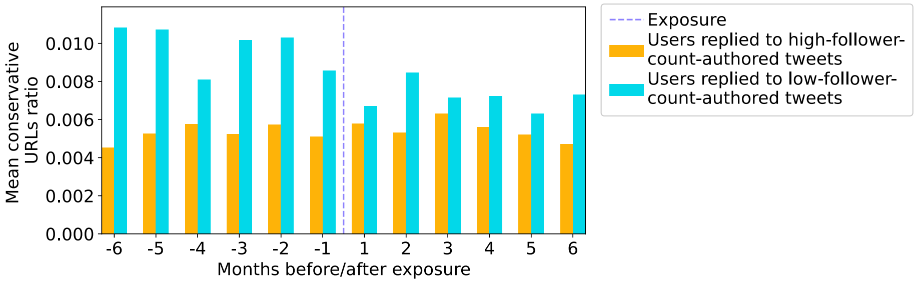


**Fig. 14** CDF plot of follower count

*Popular exposed users' (high-follower count) and less-popular users (low-follower count) are generally similar (they both have similar trends compared to the overall target group).* These two user groups do not behave differently in general. The high-follower count users have significant tweet count increases in the short-term (1.03 vs. 1.05**) and the long-term (367.5 vs. 403.5***). The low-follower count users are the same (0.56 vs. 0.60*** for short-term, and 143.2 vs. 159.6*** for long-term). For language features, we do not observe differences from the overall target group in language sentiment change, and word usage change in emotion, swear, and conflict related words. The only difference in language lies in the usage of cognitive thinking related words, where popular users have a decrease (10.75 vs. 10.70%***), but less-popular users do not have this change (10.77 vs. 10.74%, $P = 0.23$).

These two groups of users have a difference in terms of long-term liberal leaning URLs sharing, where less-popular users have a decrease (0.016 vs. 0.014***), compared with no significant change for popular users (0.019 vs. 0.018, $P = 0.08$). Their conservative media URLs sharing has no significant change, which is the same as RQ1's result.

**Fig. 15** Conservative URLs sharing comparison for users exposed to tweets posted by high and low follower authors



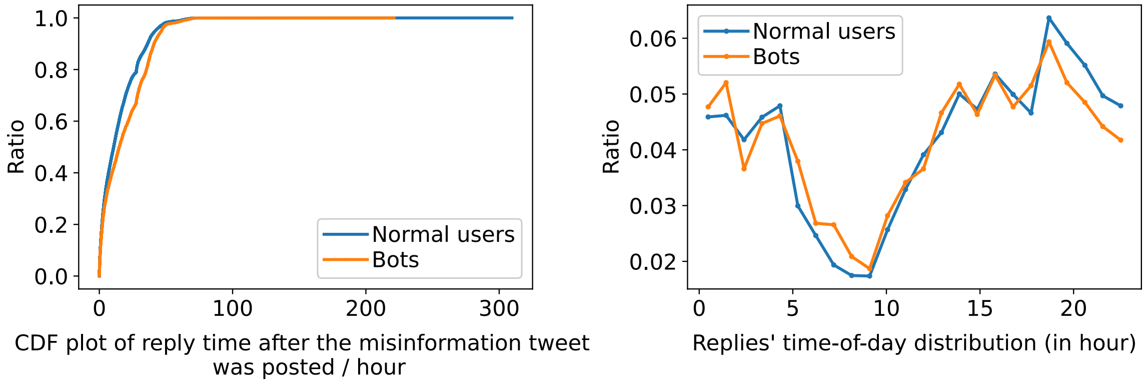


**Fig. 16** Bots and normal users' reply timing comparison

*Users exposed to high-follower-count-authored and low-follower-count-authored misinformation tweets are similar in activity and sentiment change, but differ in specific word categories' usage change.* These two user groups do not behave differently in activity and sentiment, either. Users exposed to high-follower authors' tweets have a significant increase in tweeting frequency for short term (0.78 vs. 0.80***), and long term (269.6 vs. 293.3***). Users exposed to high-follower authors' tweets have a similar trend (1.02 vs. 1.11*** for short-term, and 345.7 vs. 389.4*** for long-term). We also did not observe significant change in language sentiment.

These groups of users have a difference in terms of long-term conservative leaning URLs sharing, where users exposed to low-follower-count-authored tweets decrease (0.01 vs. 0.007**), compared with users exposed to high-follower-count-authored tweets (0.005 vs. 0.006, $P = 0.43$), which is shown in Fig. 15. Their liberal leaning media URLs sharing decreases, which is the same as RQ1's result.

Their word usage differs a lot. For users exposed to misinformation tweets posted by popular authors, their usage of cognitive thinking related words (10.77 vs. 10.71%***), emotional words (2.12 vs. 2.09%*), swear words (0.58 vs. 0.63%***) and conflict related words (0.54 vs. 0.56%***) all have the same change as the overall target group. For users exposed to tweets posted by less-popular authors, their word usages in all these categories do not have a significant change. This suggests that "influencers" potentially help more with misinformation dissemination, i.e., users are more susceptible to misinformation posted by these "influencers".

## 4.5 RQ5: How do bots behave compared with normal users in terms of their replying behavior?

We identified 3674 potential bots in our datasets. To understand bots' role in misinformation exposure, we studied their behavioral characteristics regarding the replies. As social bots' general behavior has been well studied (Varol et al 2017; Shao et al 2018), we only focused on their replying related behavior towards misinformation. Since bots' are usually managed by automation software, we examined their reply timing to explore their temporal difference with normal users. We looked at the bots' reply time lag (how long it took to reply a tweet after it was posted) and the time-of-day distribution. To our surprise, bots' replying timing is similar to normal users. As shown in Fig. 16, their reply time-lag-distribution is generally similar, and bots take a little longer to reply to tweets.

We then analyzed the bots' reply texts to explore the language characteristics. As shown in Fig. 17, their reply length (in words) is also similiar to regular users, where more bots posted shorter replies. To explore the content, we applied Short Text Topic Modeling (Qiang et al 2018) to extract the abstract topic of the bots' and normal users' replies. We used 10 topics in our case because it achieved the best coherence

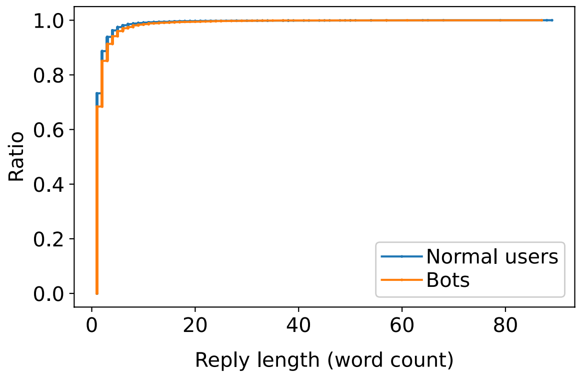


**Fig. 17** CDF plot of reply length comparison (Removed top 30 count of each group to make figure more readable)

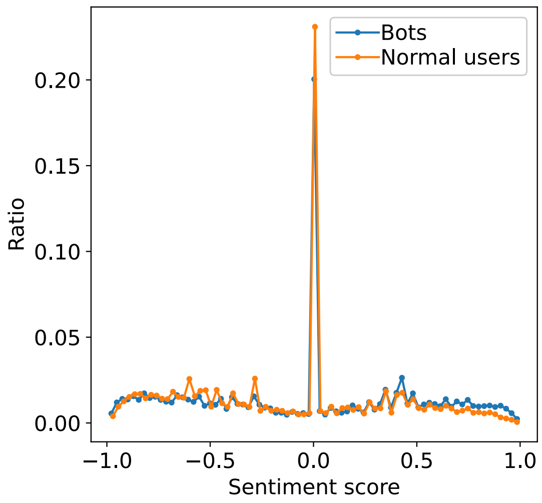


**Fig. 18** Sentiment score probability distribution for replies about US election

score. By manually examining the top words of each topic, we group the general topics into 4 categories: replies about US president election, replies about US non-election politics, replies about COVID-19, and replies about Brexit. Then we group the reply tweets into the 4 categories and calculate the sentiment score distribution in each category for bots and normal users. We find no significant difference in terms of the sentiment scores between bots and normal users. For instance, as shown in Fig. 18, their sentiment distributions are generally similar except a little more extreme language (score close to ± 1) for bots. These findings show that there is not much difference between bots and normal users when they replied to misinformation tweets, which aligns with the related work on bots' little contribution to misinformation consumption (González-Bailón and De Domenico 2021).

# 5 Concluding discussion

## 5.1 Summary and implications

This work investigates the behavior of Twitter users before and after being exposed (replied) to misinformation tweets. Our analysis reveals that users' tweet count significantly increases and liberal media URLs sharing decreases after exposure, as well as the decrease of emotional words and increase of swear and conflict related words. Meanwhile, we do not find significant change in users' sentiment. Through language distance analysis, we find that different user groups (target group and baseline group) were already different before their respective exposure, which means misinformation may attract a different group of users compared with true information. We also find that users with different political leaning change differently for some of their behaviors, which partially echoes the work (Ognyanova et al 2020) about differences in fake news consumption among different ideological subgroups. We also find that multi-exposure users have weaker tweet frequency changes than those who are only exposed once, and they are more "unstable", and are already at a high-activity level before exposure. Another finding is that the behavioral changes of popular exposed users and less-popular users are generally similar, and this finding partially holds for the users who are exposed to misinformation tweets authored by popular users and less-popular users, where we find that users exposed to misinformation posted by popular authors change their language more. We also investigate the bots' replying behavior and find there is not much difference between bots and normal users. This may help strengthen the findings in related work (González-Bailón and De Domenico 2021) about bots not being more visible than human accounts during contentious political events.

Our work reveals the positive correlation between users' behavioral changes and exposure to misinformation tweets, which can potentially encourage more research investigating the misinformation's impact on specific user behaviors. The findings on emotion, swear and conflict related word usage change provide a clue on misinformation's impact on users' language expression. Correlation between misinformation spread and political polarization has been found (Ribeiro et al 2017), and our results provide a behavioral angle on this direction, and could encourage more work on building the connection between specific behaviors and user psychological status, to further understand misinformation's impact. We also observe that both target group and baseline 2 group users decrease their word usage reflecting cognitive thinking, which is a concerning phenomenon because it has been found that lazying thinking drives the susceptibility to fake news (Pennycook and Rand 2019), and a general decrease of cognitive thinking may worsen the situation. In addition, the observation in RQ2 provides a characterization on partisan users, and reveals different behavioral change of the liberal and conservative users, especially sharing URLs from the opposite side's media. This could potentially raise more questions in understanding partisanship's role in misinformation exposure.

Our work also has important implications for social platform designers. Misinformation does not affect all users equally, and different types of users have different behavioral changes. Our second and third research questions give a closer look at these groups of interest. We also find that exposed users' behavioral changes are similar no matter how popular they are, and the language change does differ when exposed to misinformation posted by users with different popularity. These insights tell that all users could be the potential targets of misinformation, and "influencers" could be more problematic when spreading misinformation. These findings provide the social platform designers with empirical evidence and guidance, to improve the design to combat misinformation. For example, the platform should take care of all users when designing mitigation strategies because all of them could be victims no matter how popular they are. Meantime, The Key Opinion Leaders should receive more attention because they are potentially more impactful when spreading misinformation.

## 5.2 Limitations and future work

Our work does not prove causality, i.e., while we observe significant changes in behavior before and after exposure to misinformation, we cannot definitively attribute this correlation to being primarily or even exclusively caused by the exposure. Although we built the baseline groups to eliminate some factors such as user personalities (baseline 1) and the entity of the tweet (baseline 2), there may be other unforeseen factors that cause these changes. For example, long-term tweet counts may also have risen because users spend more time on Twitter. A future direction could be exploring if causality exists by careful experimental design.

Another limitation lies with respect to the dataset. First, we only collected "source" misinformation from PolitiFact, which is a small amount of misinformation and most of them are related to politics. A possible future direction is to collect more categories of misinformation (technology, business, etc) and study if changes in users' behavior are different for different types of misinformation. Second, due to Twint's limitation, this work is not able to access several critical sources of archival Twitter data including both user retweets and favorites during the necessary time period for the majority of the users. These interactions can be a good and important source to study users' attitude and preference after the exposure. As it takes different efforts to reply, retweet and favorite a tweet, a future direction is to expand the "exposure" to retweets and favorites, and compare the differences of the behavior change of users with different types of exposure. Third, the baseline 1 was generated by averaging 5 other exposures because we didn't know if other exposed tweets are misinformation, which might cause the effect of "flattening" the behavioral changes.

Our sentiment analysis also has limitation. As we used VADER which is a widely used sentiment analysis model, it might not perform well on some specific scenarios. In the context of election or COVID, some "neutral" words can have different meanings from normal use cases. For instance, a user mentioned words "Fauci" and "secret" towards a COVID vaccine related misinformation tweet, which is likely to be negative. However, the model will label it as neutral. Context-aware sentiment analysis in these specific cases is needed.

Although most users' exposure time is before or shortly after the fact-checking articles were posted, there are still a few users who may know the fact when replying to the misinformation tweets. It is also possible that some users are more resilient to misinformation than others and disagree with the content. It has been shown that users show more falsehood awareness (Jiang and Wilson 2018; Jiang et al 2020) in their replies to falser misinformation. In this work, we do not differentiate the users who agree or disagree with the misinformation. Analyzing the behavioral changes of users with different attitudes could be another fruitful direction.

This work aims to study general misinformation's impact on users. When collecting data from PolitiFact, we didn't differentiate the authenticity levels labelled by the experts (pants on fire, false, mostly false). A future direction on this can be studying if users' behavior change differently after exposure to misinformation with different authenticity level. Furthermore, as it has been shown that there are also different strategies used by misinformation (Volkova and Jang 2018; Appling et al 2015), understanding the effectiveness of different strategies and authenticity is important and useful for fighting misinformation, so that specific mitigation methods can be designed and applied accordingly.

In addition, although this work has focused on "first order" impact between the misinformation tweets and exposed users, this work may also raise the question whether impacted users also impact their friends and followers through their retweets, replies and mentions, i.e., the "second order" impact.

**Funding** This study was partially funded by the Google Cloud Research Credits program with the credit# 204357824, and Summer Graduate School Fellowship from University of Colorado Boulder.

## Declarations

**Conflict of interest** All authors declare that they have no conflict of interest.

**Ethics approval** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

Alba D (2020) On facebook, misinformation is more popular now than in 2016. https://www.nytimes.com/2020/10/12/technology/on-facebook-misinformation-is-more-popular-now-than-in-2016.html

Aldayel A, Magdy W (2019) Assessing sentiment of the expressed stance on social media. In: International conference on social informatics, Springer, pp 277–286

Althoff T, Clark K, Leskovec J (2016) Large-scale analysis of counseling conversations: an application of natural language processing to mental health. Trans Assoc Comput Linguist. 5:996

Appling DS, Briscoe EJ, Hutto CJ (2015) Discriminative models for predicting deception strategies. In: International conference on World Wide Web

Benevenuto F, Rodrigues T, Cha M, et al (2009) Characterizing user behavior in online social networks. In: ACM SIGCOMM conference on internet measurement

Bessi A, Coletto M, Davidescu GA et al (2015) Science vs conspiracy: collective narratives in the age of misinformation. PloS ONE 2:788

Bianco R (2021) Study finds misinformation still driving vaccine hesitancy. https://www.10news.com/about/10news-team/study-finds-misinformation-still-driving-vaccine-hesitancy

Boyd RL, Ashokkumar A, Seraj S et al (2022) The development and psychometric properties of liwc-22. University of Texas at Austin, Austin

Ciampaglia GL, Shiralkar P, Rocha LM et al (2015) Computational fact checking from knowledge networks. PLoS ONE 2:1400

Del Vicario M, Bessi A, Zollo F et al (2016) The spreading of misinformation online. Proc Natl Acad Sci U S A 1:799

Dutta U, Hanscom R, Zhang JS et al (2021) Analyzing twitter users' behavior before and after contact by Russia's internet research agency. Proc ACM Hum Comput Interact 2:666

Feng S, Banerjee R, Choi Y (2012) Syntactic stylometry for deception detection. In: Annual meeting of the association for computational linguistics

Fuglede B, Topsoe F (2004) Jensen-shannon divergence and hilbert space embedding. In: International symposium on information theory (ISIT), IEEE

González-Bailón S, De Domenico M (2021) Bots are less central than verified accounts during contentious political events. Proc Natl Acad Sci 118(11):550

Hassan N, Arslan F, Li C, et al (2017) Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In: International conference on knowledge discovery and data mining (SIGKDD)

Hessel J, Tan C, Lee L (2016) Science, askscience, and badscience: on the coexistence of highly related communities. In: International AAAI conference on web and social media

Holme P, Rocha LE (2019) Impact of misinformation in temporal network epidemiology. Netw Sci 2:447

Horne B, Adali S (2017) This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Proceedings of the international AAAI conference on web and social media, pp 759–766

Hutto C, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: International AAAI conference on web and social media

Javed RT, Shuja ME, Usama M, et al (2020) A first look at covid-19 messages on whatsapp in Pakistan. In: 2020 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), https://doi.org/10.1109/ASONAM49781.2020.9381360

Jiang S, Wilson C (2018) Linguistic signals under misinformation and Fact-Checking: Evidence from user comments on social media. Proc ACM Hum-Comput Interact 2:1–23

Jiang S, Metzger M, Flanagin A, et al (2020) Modeling and measuring expressed (dis) belief in (mis) information. In: Proceedings of the international AAAI conference on web and social media, pp 315–326

Jin L, Chen Y, Wang T et al (2013) Understanding user behavior in online social networks: a survey. IEEE Commun Mag 3:4470

Klepper D (2021) Defense for some capitol rioters: election misinformation. https://apnews.com/article/dc-wire-donald-trump-health-coronavirus-pandemic-election-2020-b7e929bb8d49b77d0922eae7ad3794b7

Lemon J (2021) Dominic pezzola is latest capitol rioter to blast trump for misleading supporters. https://www.newsweek.com/dominic-pezzola-latest-capitol-rioter-blast-trump-misleading-supporters-1568351

Loomba S, de Figueiredo A, Piatek SJ et al (2021) Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2:1550

Luceri L, Deb A, Badawy A, et al (2019) Red bots do it better: comparative analysis of social bot partisan behavior. In: Companion proceedings of the 2019 World Wide Web conference, pp 1007–1012

Ma J, Gao W, Wong KF (2018) Rumor detection on Twitter with tree-structured recursive neural networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp 1980–1989, https://doi.org/10.18653/v1/P18-1184

Mocanu D, Rossi L, Zhang Q et al (2015) Collective attention in the age of (mis) information. Comput Hum Behav 3:1778

Mohammad SM, Sobhani P, Kiritchenko S (2017) Stance and sentiment in tweets. ACM Trans Internet Technol 2:668

Mustafaraj E, Metaxas PT (2017) The fake news spreading plague: was it preventable? In: ACM on web science conference

Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Political Behav 4:800

Ognyanova K, Lazer D, Robertson RE, et al (2020) Misinformation in action: fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School Misinformation Review

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

Pérez-Rosas V, Kleinberg B, Lefevre A et al (2017) Automatic detection of fake news. arXiv preprint arXiv:1708.07104

Picchi A (2018) Fake news: Twitter still flooded with sham accounts. https://www.cbsnews.com/news/fake-news-twitter-still-flooded-with-sham-accounts/

Qiang J, Li Y, Yuan Y, et al (2018) Sttm: a tool for short text topic modeling. arXiv preprint arXiv:1808.02215

Ribeiro MH, Calais PH, Almeida VA, et al (2017) "everything i disagree with is# fakenews": correlating political polarization and spread of misinformation. arXiv preprint arXiv:1706.05924

Shao C, Ciampaglia GL, Varol O, et al (2017) The spread of fake news by social bots. arXiv preprint arXiv:1707.07592 96:104

Shao C, Ciampaglia GL, Varol O et al (2018) The spread of low-credibility content by social bots. Nat Commun 9(1):1–9

Shu K, Wang S, Liu H (2018) Understanding user profiles on social media for fake news detection. In: Conference on multimedia information processing and retrieval (MIPR), IEEE

Shu K, Wang S, Liu H (2019) Beyond news contents: the role of social context for fake news detection. In: International conference on web search and data mining

Tambuscio M, Ruffo G, Flammini A, et al (2015) Fact-checking effect on viral hoaxes: a model of misinformation spread in social networks. In: International conference on world wide web

Varol O, Ferrara E, Davis C, et al (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the international AAAI conference on web and social media

Vicario MD, Quattrociocchi W, Scala A et al (2019) Polarization and fake news: early warning of potential misinformation targets. ACM Trans Web 2:174

Volkova S, Jang JY (2018) Misleading or falsification: inferring deceptive strategies and types in online news and social media. In: Companion proceedings of the the web conference 2018

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 3:500

Wang Y, Han R, Lehman T, et al (2021) Analyzing behavioral changes of twitter users after exposure to misinformation. In: Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining, pp 591–598

Wu K, Yang S, Zhu KQ (2015) False rumors detection on sina weibo by propagation structures. In: International conference on data engineering

Zhou X, Jain A, Phoha VV et al (2020) Fake news early detection: A theory-driven model. Res Pract Digit Threat

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.