2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Analyzing Behavioral Changes of Twitter Users After Exposure to Misinformation

Yichen Wang, Richard Han, Tamara Lehman, Qin Lv, and Shivakant Mishra
University of Colorado Boulder, Boulder, CO, USA
{yichen.wang, richard.han, tamara.lehman, qin.lv, shivakaht.mishra}@colorado.edu

***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 behavior of exposed users to determine whether the frequency of the tweets they posted, or the sentiment of their tweets underwent any significant change. Our results indicate 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 two specific user groups, multi-exposure and extreme change groups, which were potentially highly impacted. Finally, we study if the changes in the behavior of the users after exposure to misinformation tweets vary based on the number of their followers or the number of followers of the tweet authors, and find that their behavioral changes are all similar.**

***Index Terms*—Misinformation, Fake News, Twitter, User Behavior**

I. 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 [1], [2]. As another example, misinformation is still driving people's vaccine hesitancy, especially during the COVID-19 pandemic [3]. 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 [4]–[6]. To make it more amenable to the masses, academics have also proposed mechanisms to automate the fact-checking process [7], [8].

To study misinformation's impact, researchers have investigated the spread pattern of misinformation [9], [10], its negative effect on users' beliefs [11]–[13], and its correlation with some social phenomena such as echo chambers and polarization [14]. However, prior work has focused more on the misinformation's general social effect, and very little 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 related misinformation on users' vaccine intent [15], 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 he/she replies to a tweet carrying misinformation (*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: Does the change in behavior of users who reply to *multiple* misinformation tweets differ from other users?
- RQ3: What are the characteristics of the users who undergo extreme behavioral change after being exposed to misinformation tweets?
- 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?

To identify misinformation tweets, we first obtained fact-checked misinformation excerpts from the well-known fact-

ASONAM '21, November 8-11, 2021, Virtual Event, Netherlands

ACM ISBN 978-1-4503-9128-3/21/11...$15.00
https://doi.org/10.1145/3487351.3492718

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 count, mean sentiment score of tweets, and language usage distance. 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). 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 the number and frequency of tweets that users send after being exposed to misinformation tweets, both in the short and long terms. We do not observe such changes in the baseline user groups, indicating that there is a positive correlation between increase in count/frequency of tweets and exposure to misinformation tweets.
- We do not find any significant change in user's overall tweet sentiments after being exposed to misinformation tweets. Using language distance analysis, we show that baseline users already had different language characteristics than the exposed users even before the exposure.
- 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 significant when compared with the users exposed to a single misinformation tweet, and that these users were already on a high activity level before the exposures.
- We examine the group of users who had extreme behavioral changes and find that users with extreme changes of tweet count do not overlap with the group of users with extreme change of sentiment. Further, users with extreme changes in the short term do not overlap with the users with extreme changes in the long-term.
- We find that exposed users with high and low follower counts exhibit similar behavioral change (significantly increase tweeting frequency). Further, we find users exposed to misinformation tweets authored by users with high and low follower counts also undergo similar behavioral change, while the users exposed to low-follower-count-authored tweets generally post more tweets.

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 Section II, we describe the background and related work on misinformation on social media. In Section III, we present and explain the methodologies used to create the dataset and the user behavioral features under study. In Section IV, we present the results of the analysis of the research questions. Finally, in Section V, we conclude the work and discuss its implications, limitations, and possible future directions.

II. Related Work

Researchers have studied users and content on online social platforms extensively [16]–[18], and it has been shown that online social media has become a major source of misinformation [19]–[21]. A significant body of work has investigated the spread and detection of misinformation. Mustafaraj et al. described the spread process of fake news [22]. The diffusion process is also modeled by Tambusc et al. [23]. 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 [10]. Vicario et al. studied the conspiracy news spreading on Facebook and found selective exposure is the primary driver of the diffusion [9]. 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 [7], [8]. Researchers have investigated automatic detection through content style [4], [24], [25], user profile [5], and information propagation [6], [26], [27]. Our work differs from this body of work in that we focus on the users to understand what and how their behavior changed after being exposed to misinformation.

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 [12], [13]. Scientists have also shown that misbeliefs can persist after being exposed to misinformation [11]. Loomba et al. focused on COVID vaccine related misinformation and found that users' vaccine intent decreased after exposure via qualitative analysis [15]. 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 [28]. Researchers have investigated the correlation between misinformation and society polarization [14], [29]. Holme et al. also studied the influence on social network structure via

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

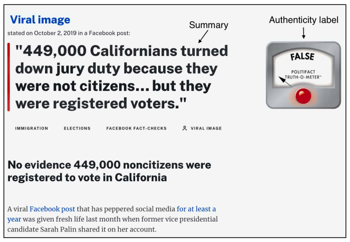
simulations [30]. In contrast, our work investigates the specific behavioral differences of users before and after the exposure to misinformation. To the best of our knowledge, this is the first work performing a large-scale, quantitative analysis on users' behavioral change after being exposed to a broad range of misinformation.

III. Methodology

A. Data Collection

1) Collecting Misinformation Tweets and the Exposed Users' 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 from PolitiFact as our "seed" to find the corresponding tweets. Since PolitiFact does not work on tweets, we crawled all the fact-checked Facebook text-only and viral image posts which are labelled as "pants on fire", "false" or "mostly false" from May 18, 2013 until Jan 31, 2021 (most of them are during 2018 to 2021). We crawled Facebook posts as it is also a social network platform and posts have a high probability of showing up on Twitter. Although PolitiFact also debunks politicians' and celebrities' claims, it turned out to be harder to directly search for them on Twitter. Figure 1 is an example of a fact-checked Facebook post on PolitiFact. For each debunked news post, we used the provided summary as the search term to search for the corresponding tweets on Twitter. To avoid unrelated results, we disregarded the posts whose summary was less than seven words. From the search response, we extracted the top-five tweets ranked by reply count. We removed the tweets that originated from fact-checking organizations, or that included any keyword regarding its veracity, e.g. "conspiracy theory", "debunk", and "fake news". We crawled 1,119 debunked news posts from PolitiFact and we found 442 of them on Twitter. From the search results, we were able to collect 529 tweets with misinformation content. Figure 2 shows an example of a tweet that we studied. 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.

Fig. 1: A PolitiFact article debunking a Facebook post



Viral image

stated on October 2, 2019 in a Facebook post:

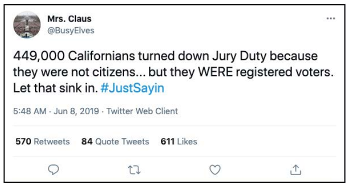
"449,000 Californians turned down jury duty because they were not citizens... but they were registered voters."

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 sample tweet containing the misinformation



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

For each of the collected tweets, we identified all the users who replied to it, which resulted in 25,619 users. 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.

In order to ensure the users under the purview of this study were legitimate users, we used Botometer [31] 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 more than 0.5 were removed. There were 372 misinformation tweets with 21,071 replied users for target group after potential bot removal.

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.

2) Generating the Baseline User Tweets Dataset: In order to ensure that observed user behavior was sufficiently different from the general Twitter population, we built two user groups as baseline groups. Both groups were used as the baselines for the short-term analysis (24 hours) and one of them was used for the long-term analysis (6 months). For the first baseline group, we used the target group to understand the behavioral change when the same users were 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 analyses on this baseline show the average behavior of the 5 exposures. We selected 5 exposures because we were not able to confirm if other exposures were to true news tweets, so we averaged multiple exposures to eliminate it. 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 analysis.

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 [32]. Then, we collected all the recent tweets from known true news sources [33] 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. Table I shows 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.

Retweets and favorite tweets were not utilized within this work, as their presence could be unevenly distributed due to a time-sensitive constraint. We used Twint [34], 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.

TABLE I: Number of users for the analysis

User group	Analysis type	No. of users
Target group (same as baseline 1)	Short-term	21,071
	Long-term	11,585
Baseline 2	Short-term	19,357
	Long-term	5,970

B. Features

We analyzed the before and after user behavior through three specific metrics: average tweet count, average sentiment score, and language usage distance. All features were studied hourly (short-term) and monthly (long-term).

Tweet Count 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. We excluded favorite tweets, retweets of tweets authored by another Twitter account, or other Twitter activity from the tweet count because of the limitations mentioned in the data collection section.

Sentiment Score was calculated using VADER, a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiment expressed in social media [35]. 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. These values were then averaged to form the hourly or monthly *Sentiment Score* for each user.

Language Usage Distance was used to evaluate the difference of language between two groups of tweets text. Similar to prior work [36], [37], we adopted the Jensen-Shannon Divergence [38] 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 step.

We used dependent sample t-test to assess the statistical significance of the results for both *Tweet Count* and *Sentiment Score*. We used this type of test because the before and after samples are not independent of each other. We aggregated the users' hourly/monthly feature before and after the exposure to conduct the tests.

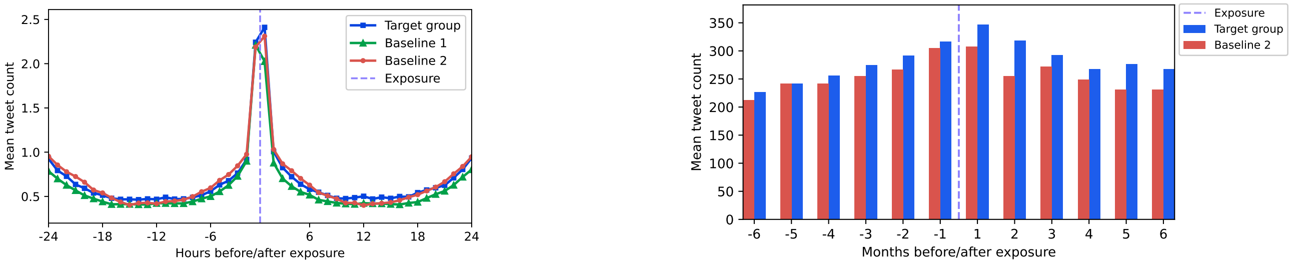
IV. Results

A. RQ1: Do the users who reply to misinformation tweets exhibit a change in their behavior after the exposure?

The target group significantly increased tweeting frequency following their exposure to the misinformation tweets in both short and long term, compared with the baseline users. As shown in Fig. 3a, in the short-term analysis all three groups' tweeting frequency had a 24-hour periodic change. The target group's tweeting frequency increased significantly (0.66 vs. 0.68, P=5.7×10^{-12}) during the 24-hour period after the exposure, while baseline 1 group's decreased significantly (0.60 vs. 0.59, P=4.1×10^{-5}). 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 showed the similar pattern. Due to space limitations, we only report the 24-hour analysis.

The long-term analysis had a similar pattern, as shown in Fig. 3b. Although monthly tweet count increased for both groups, the target group's tweet count increased more significantly. The baseline 2 group's change was significant (253.5 vs. 257.6, P=0.013), but the change and significance level is much weaker than that of the target group (267,7 vs. 294.9, P=2.6×10^{-132}).

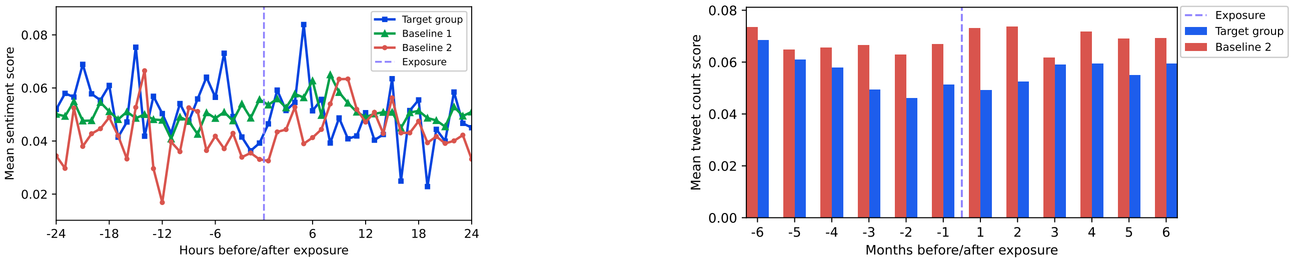
The target group users did not change sentiment significantly in neither short nor long term. Sentiment score of all the 3 groups in the short-term did not change significantly (Fig. 4a). As shown in Fig. 4b, the target group's sentiment



(a) Short-term: hourly tweet count

(b) Long-term: monthly tweet count

Fig. 3: Average hourly (left) and monthly (right) tweet count



(a) Short-term: hourly sentiment score

(b) Long-term: monthly sentiment score

Fig. 4: Average hourly (left) and monthly (right) sentiment score

score didn't change significantly in long-term either (0.056 vs. 0.056, P=0.85), while baseline 2 group's sentiment score had a little increase (0.067 vs. 0.070, P=1.3×10^{-5}). 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 negative or positive language [39], [40], which would not change the average sentiment score significantly.

To understand this 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. 5, the language distance of the target group with the baseline group is stable before and after the exposure, and there is a slight increase starting from the fourth month after exposure. This observation indicates that the target and the baseline 2 group users already have different language characteristics even before their respective exposure and that this difference does not change much after the exposure. This indicates that the misinformation and true information tweets attract users with different characteristics.

B. RQ2: *Does the change in behavior of users who reply to multiple misinformation tweets differ from other users?*

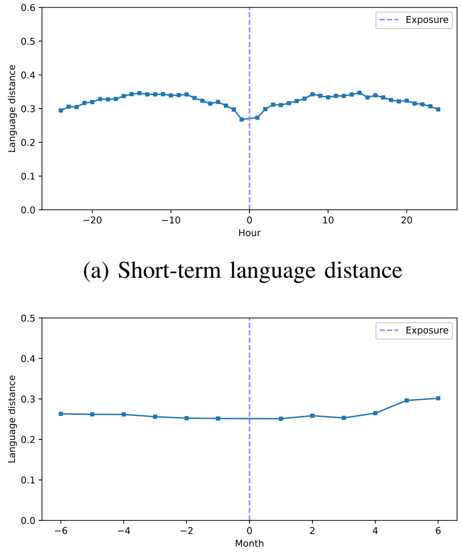
Multi-exposure users show a significant change of tweet count in long-term but not in short-term, and the change is weaker than that of other users. We consider multi-exposure users to be the ones who replied to at least two misinformation tweets. Although users may reply to other misinformation tweets,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

As shown in Fig. 6, the tweet count for multi-exposure users did not have significant change in the short-term (1.14 vs. 1.17, P=0.37), while other users' (single-exposure) increase was statistically significant (0.64 vs. 0.67, P=7.3×10^{-12}). In the long-term, multi-exposure users still had an increased tweet count (465.0 vs. 523.0, P=7.3×10^{-13}), but its significance level is weaker than that of the single-exposure users (261.2 vs. 287.3, P=1.3×10^{-121}). As mentioned in RQ1's result, we did not observe significant sentiment change .

From the comparison between the multi-exposure and single-exposure groups, it is shown that the multi-exposure group generally posts more tweets (Fig. 6) with more volatile sentiment (Fig. 7), and 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 were already on a "high-level mood" and their change was not as significant as that of the single-exposure users, who were rising from a relatively lower level.

C. RQ3: *What are the characteristics of the users who undergo extreme behavioral change after being exposed to misinformation tweets?*

Another interesting angle is to study the users who changed their behavior the most after the exposure. To separate this group of users, we calculate the tweet count increase and absolute sentiment score change from 1 hour/month before to 1 hour/month after the exposure. The users who had



(a) Short-term language distance

(b) Long-term language distance

Fig. 5: Language distance between target and baseline 2 group

the top 5% tweet count increase or sentiment change were selected, respectively. This resulted in 1,007 and 939 users who had extreme tweet count increase for the first hour/month, respectively. The the first hour/month extreme sentiment score change group had 1,054 and 941 users, respectively.

Extreme tweet count increase and extreme sentiment score change do not align. We first examined if the users having extreme tweet count increase and those having extreme sentiment change overlap. There were only 7 and 8 overlapped users for the first hour/month, respectively. This means that after being exposed to misinformation, “furiously” posting tweets didn’t occur together with sharp change of sentiment.

Short-term and long-term change do not align. We also compared the tweeting frequency between the short and long-term among the same extreme-change users. Users who increased tweet count a lot in the first hour didn’t show a large increase in the following months. Similarly, the users who increased tweeting frequency a lot in the first month didn’t show the same level of increase in the first several hours. There were 146 and 62 overlapped users for extreme tweet count and sentiment change, respectively. Fig. 8 visualizes the overlap across different extreme-change user groups.

D. 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. Fig.9 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.

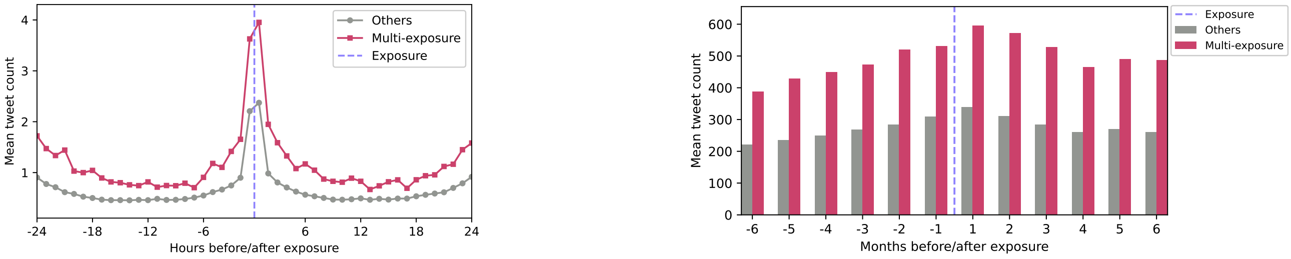
Popular exposed users’ (high-follower count) and less-popular users (low-follower count) both increased their tweeting frequency. These two user groups didn’t behave differently. The high-follower count users did have significant tweet count increases in short-term (1.03 vs. 1.05, P=0.006) and long-term (367.5 vs. 403.5, $P=4.3\times10^{-86}$). The low-follower count users were similar (0.56 vs. 0.60, $P=4.3\times10^{-12}$ for short-term, and 143.2 vs. 159.6, $P=8.2\times10^{-58}$ for long-term). As mentioned in RQ1’s result, we did not observe significant sentiment change.

Users exposed to high-follower-count-authored and low-follower-count-authored misinformation tweets both increased their tweeting frequency. These two user groups didn’t behave differently, either. Users exposed to high-follower authors’ tweets had a significant increase in tweeting frequency for short term (0.78 vs. 0.80, $P=1.03\times10^{-6}$), and long term (269.6 vs. 293.3, $P=8.1\times10^{-85}$). Users exposed to high-follower authors’ tweets were similar (1.02 vs. 1.11, $P=3.6\times10^{-7}$ for short-term, and 345.7 vs. 389.4, $P=9.1\times10^{-17}$ for long-term). Fig.10 shows the long-term change and the users exposed to low-follower-count-authored tweets generally post more tweets. As mentioned in RQ1’s result, we did not observe significant sentiment change.

V. Concluding Discussion

This paper investigates the behavior of Twitter users before and after being exposed (replied) to misinformation tweets. Our analysis reveals that users’ tweet count significantly increases after exposure in both short and long-term. We do not find significant change in users’ sentiment score. Through language distance analysis, we find that different user groups (target group and baseline group) are already different before their respective exposure. We also find that users who are exposed to more than one misinformation tweet have weaker changes than those who are only exposed to one, and find these multi-exposure users are already at a high-activity level before exposure. For users who have extreme behavioral changes, we find that their tweet count increases and sentiment score changes do not align. The short-term and long-term change do not align either. Another finding is that popular exposed users and less-popular users both increased their tweeting frequency, and this finding also holds for the users that are exposed to misinformation tweets authored by high-follower count users and low-follower count users, while the users exposed to low-follower-count-authored tweets generally post more tweets.

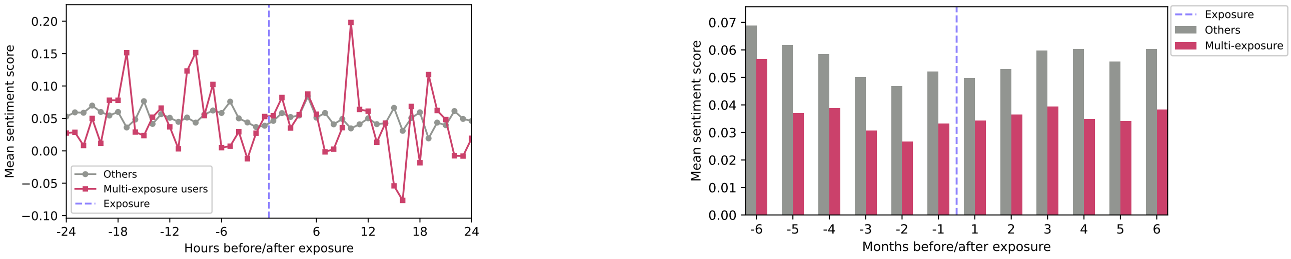
Implications. Our work reveals the positive correlation between users’ tweeting frequency increase and exposure to misinformation tweets, which can potentially encourage more



(a) Short-term: hourly tweet count

(b) Long-term: monthly tweet count

Fig. 6: Average hourly (left) and monthly (right) tweet count for multi-exposure users.



(a) Short-term: hourly sentiment score

(b) Long-term: monthly sentiment score

Fig. 7: Average hourly (left) and monthly (right) sentiment score for multi-exposure users.

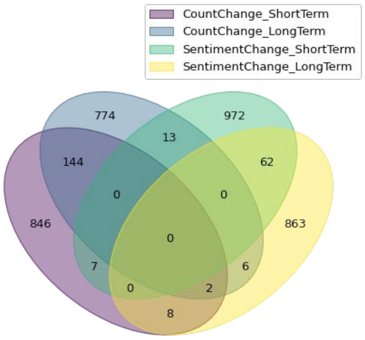


Fig. 8: Number of users in the extreme-change user groups.

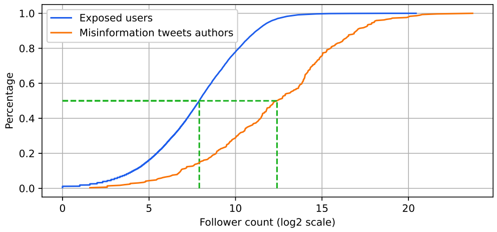


Fig. 9: CDF plot of follower count

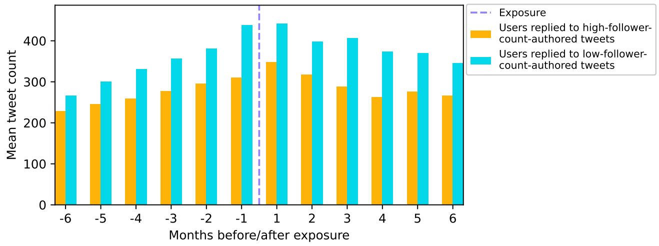


Fig. 10: Average hourly tweet count for exposed high-follower and low-follower count users

research investigating the misinformation's impact on specific user behaviors. Our work also has important implications for social platform designers and moderators. Misinformation does not affect all users equally and only a small number of users exhibit significant behavioral changes. Our second and third research questions give a closer look at these groups of interest. We also find that the behavioral changes are similar for exposed users and misinformation tweets authors' with different follower count. These insights tell that all users could be potential target or disseminator of misinformation, which means the platform moderators should take care of all users when designing misinformation mitigation strategies.

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.

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 Tweepy'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 behavioural changes.

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 [41], [42], 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 of whether impacted users also impact their friends and followers through their retweets, replies and mentions, i.e., the "second order" impact.

References

[1] D. Klepper, "Defense for some capitol rioters: election misinformation," May 2021. [Online]. Available: https://apnews.com/article/dc-wire-donald-trump-health-coronavirus-pandemic-election-2020-b7e929bb8d49b77d0922eae7ad3794b7

[2] J. Lemon, "Dominic pezzola is latest capitol rioter to blast trump for misleading supporters," Feb 2021. [Online]. Available: https://www.newsweek.com/dominic-pezzola-latest-capitol-rioter-blast-trump-misleading-supporters-1568351

[3] R. Bianco, "Study finds misinformation still driving vaccine hesitancy," Jun 2021. [Online]. Available: https://www.10news.com/about/10news-team/study-finds-misinformation-still-driving-vaccine-hesitancy

[4] V. Pérez-Rosas *et al.*, "Automatic detection of fake news," Aug. 2017.

[5] K. Shu *et al.*, "Understanding user profiles on social media for fake news detection," in *Conf. on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018.

[6] ——, "Beyond news contents: The role of social context for fake news detection," in *Intl. Conf. on Web Search and Data Mining*, 2019.

[7] N. Hassan *et al.*, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2017.

[8] G. L. Ciampaglia *et al.*, "Computational fact checking from knowledge networks," *PLoS One*, 2015.

[9] M. Del Vicario *et al.*, "The spreading of misinformation online," *Proc. Natl. Acad. Sci. U. S. A.*, 2016.

[10] S. Vosoughi *et al.*, "The spread of true and false news online," *Science*, 2018.

[11] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, 2010.

[12] A. Bessi *et al.*, "Science vs conspiracy: Collective narratives in the age of misinformation," *PloS one*, 2015.

[13] D. Mocanu *et al.*, "Collective attention in the age of (mis) information," *Computers in Human Behavior*, 2015.

[14] M. H. Ribeiro *et al.*, ""Everything I Disagree With is #FakeNews—: Correlating political polarization and spread of misinformation"," 2017.

[15] S. Loomba *et al.*, "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA," *Nature human behaviour*, 2021.

[16] F. Benevenuto *et al.*, "Characterizing user behavior in online social networks," in *ACM SIGCOMM Conference on Internet Measurement*, 2009.

[17] L. Jin *et al.*, "Understanding user behavior in online social networks: A survey," *IEEE Communications Magazine*, 2013.

[18] Y. Wang *et al.*, "Jump on the bandwagon?–characterizing bandwagon phenomenon in online nba fan communities," in *International Conference on Social Informatics*. Springer, 2020.

[19] A. Picchi, "Fake news: Twitter still flooded with sham accounts," Oct 2018. [Online]. Available: https://www.cbsnews.com/news/fake-news-twitter-still-flooded-with-sham-accounts/

[20] D. Alba, "On facebook, misinformation is more popular now than in 2016," Oct 2020. [Online]. Available: https://www.nytimes.com/2020/10/12/technology/on-facebook-misinformation-is-more-popular-now-than-in-2016.html

[21] R. T. Javed *et al.*, "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)*, 2020.

[22] E. Mustafaraj and P. T. Metaxas, "The fake news spreading plague: was it preventable?" in *ACM on web science conference*, 2017.

[23] M. Tambuscio *et al.*, "Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks," in *International Conference on World Wide Web*, 2015.

[24] S. Feng *et al.*, "Syntactic stylometry for deception detection," in *Annual Meeting of the Association for Computational Linguistics*, 2012.

[25] X. Zhou *et al.*, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, 2020.

[26] J. Ma *et al.*, "Rumor detection on twitter with tree-structured recursive neural networks." Association for Computational Linguistics, 2018.

[27] K. Wu *et al.*, "False rumors detection on sina weibo by propagation structures," in *International Conference on Data Engineering*, 2015.

[28] U. Dutta *et al.*, "Analyzing twitter users' behavior before and after contact by russia's internet research agency," *Proc. ACM Hum.-Comput. Interact.*, 2021.

[29] M. D. Vicario *et al.*, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Trans. Web*, 2019.

[30] P. Holme and L. E. Rocha, "Impact of misinformation in temporal network epidemiology," *Network Science*, 2019.

[31] "Botometer API," https://botometer.osome.iu.edu/api, accessed: 2021-05-30.

[32] "CoreNLP," https://stanfordnlp.github.io/CoreNLP/.

[33] B. D. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," Mar. 2017.

[34] "Twint," https://github.com/twintproject/twint, accessed: 2021-05-30.

[35] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *International AAAI Conference on Web and Social Media*, 2014.

[36] J. Hessel *et al.*, "Science, askscience, and badscience: On the coexistence of highly related communities," in *International AAAI Conference on Web and Social Media*, 2016.

[37] T. Althoff *et al.*, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *Trans Assoc Comput Linguist*, 2016.

[38] B. Fuglede and F. Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *Intl. Symp. on Information Theory (ISIT)*. IEEE, 2004.

[39] S. M. Mohammad *et al.*, "Stance and sentiment in tweets," *ACM Trans. Internet Technol.*, 2017.

[40] A. Aldayel and W. Magdy, "Assessing sentiment of the expressed stance on social media," in *International Conference on Social Informatics*. Springer, 2019, pp. 277–286.

[41] S. Volkova and J. Y. Jang, "Misleading or falsification: Inferring deceptive strategies and types in online news and social media," in *Companion Proceedings of the The Web Conference 2018*, 2018.

[42] D. S. Appling *et al.*, "Discriminative models for predicting deception strategies," in *International Conference on World Wide Web*, 2015.