Sydney Upcraft and Abby Wright

MI 355 740

Professor Peng

4 May 2023

Research Paper

Digital literacy has become increasingly important in the new age of social media and technology, as we, as a society, grow into new means of living and working via our phones, computers, tablets, etc. However, due to lack of access and issues worldwide of an expanding understanding of what digital literacy is, there are still large populations of people who do not properly know how to assess information they find on the internet. Misinformation today is a parasite that spreads through the use of technology, and with the help of digitally illiterate people, a large amount of misinformation makes its way into outlets that are readily accessible to others to see, such as social media, news outlets, and more. Thus, due to the importance of learning about how to quell issues of misinformation, the following paper studies the research question: Does the technique of flagging sensitive or incorrect content reduce the distribution of misinformation on Twitter?

A research study by the University of Alabama in the National Library of Medicine entitled "Use of bot and content flags…" looks at the flagging feature on Twitter and its connection to engagement on tweets, a tool we used to focus on our research study. Though the focus of this study is on the effects of misinformation spread and the use of flags to counteract false information regarding the pandemic, the results have helped to inspire the following study. In the study, the results concluded that "Twitter message flags negatively affect participants' perceptions of unreliable tweets." (Lanius et al., 2021). The study also states that "our results

show that merely identifying misinformation, without offering corrections, can decrease participants' opinions of the misinformation source." (Lanius et al. 2021)

Another study by Allen, Martel, and Rand posted in 2021 titled, "Birds of a feather don't fact-check each other…" focuses on the 2021 launch of Twitter's Birdwatch program. Although it has since transformed into Community Notes as of November 2022 (Coleman, 2023), the original use of Birdwatch was to identify misinformation, though in early phases, this was done on a separate site. Though the focus of the study was conducted around partisanship, the information within the study allows an understanding of the beginnings of Birdwatch by Twitter and how it has since grown into what it is now: Community Notes. This helps us to understand that the use of flagging information, on any website, not just Twitter, is still a new thing that is growing and changing as we reevaluate how to grow digital literacy. In the beginning, based on the study by Allen, Martel, and Rand, we know that originally Birdwatch was used as a way to show that a tweet was untrue, but not specify why it was not unless a user visited a separate website. It has since transformed to now allow for information given on the Twitter application.

However, on the other side of the spectrum, a study titled, "'This is fake news': Investigating the role of conformity to other users' views…" by Jonas Colliander published in the journal 'Computers in Human Behavior' found the opposite data. Colliander states that "... it thus validates recent decisions by social media companies to move away from flagging options in responses to fake news as they do not seem to be particularly effective." (Colliander, 2019, p. 208) Colliander found that "a disclaimer is not as effective as other users' comments in stopping the spread of fake news." (Colliander, 2019, p. 207) With these articles, both sides of the research show conflicting data, and with the study we conducted, our findings point to what our data displayed in terms of the effects of flagging misinformation on Twitter.

Method:

The study was conducted via survey using Qualtrics through Michigan State University, which was accessible by both mobile and computer devices. A random sample of 39 participants were recruited via the social media platform, Twitter. They could complete the survey through the link provided in a tweet from a public account. No personal information was gathered from participants, and all responses were kept anonymous. The survey was clearly labeled as voluntary, and there were no compensational benefits offered. The independent variable was the exposure to a tweet with a context flag indicating it contains sensitive or incorrect content. The dependent variable was then the subsequent distribution of misinformation on Twitter. A “context flag” is a warning label provided on a tweet that may contain sensitive or incorrect content, which is indicated by the phrase, “Readers added context they thought people might want to know”.

We chose to distribute this survey through Twitter because we wanted to perceive the share of information via actual users of the platform. All participants were questioned whether or not they use Twitter, and were prompted to either answer “yes” or “no”. In order to get better results, we did not include the responses of those who chose “no”, resulting in an updated sample of 34. Users were then randomly assigned to either the treatment group or the control group. Both groups were given two screenshots of tweets. The treatment group was shown tweets containing context flags provided by Twitter, and the control group was shown the same tweets, but without context flags. For each tweet they were asked to rate their likelihood to retweet or share, using a 5-point Likert scale of "Very Unlikely" (1) to "Very Likely" (5). In addition, the participants were asked how a context flag has the potential to impact their responses. The treatment group was also asked whether they believed the presence of the context flag impacted

their previous responses, while the control group was asked whether the addition of a flag would impact their previous responses. Responses to these final questions were split into three choices, being “yes”, “maybe”, or “no”.

These tweets were chosen based on topic and overall popularity, so we decided to include one based on pop culture, which was flagged as misinformation based on the idea of a collaboration of artists on a new song. In contrast to this, we included a tweet with a picture of the moon from the official account of Samsung Mobile, which was flagged because a Samsung phone did not actually take the picture. Since these tweets were vastly different in topic, it was less targeted towards one demographic or pool of Twitter users in order to increase external validity.

Results:

Since we filtered the reports to ensure that the survey results only included Twitter users, question one was not analyzed, since there were zero responses of “no”. After giving each participant a random assignment to a group, the treatment group ended up having 18 participants, and the control group ended up having 16 participants. To examine the data from the survey in relation to our research question, there were multiple t-tests conducted. As shown in table 1, when analyzing the responses by the treatment group, the flagged pop culture tweet resulted in a mean score of 1.28 (SD = 0.56), and the second question regarding the flagged Samsung image had a mean score of 1.17 (SD = 0.37) based on the 5-point likert scale. As shown in table 2, for the control group, the unflagged pop culture tweet resulted in a mean score of 2.25 (SD = 1.09), and the second question regarding the unflagged Samsung image had a mean score of 2.44 (SD = 0.86).

Table 1. T-test results to the flagged tweets

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
On a scale of "Very Unlikely" (1) to "Very Likely" (5), how likely are you to retweet or share the above tweet?	1.00	3.00	1.28	0.56	0.31	18
Field	Min	Max	Mean	Standard Deviation	Variance	Responses
On a scale of "Very Unlikely" (1) to "Very Likely" (5), how likely are you to retweet or share the above tweet?	1.00	2.00	1.17	0.37	0.14	18

Note. These results were taken from the responses of the treatment group.

Table 2. T-test results to the unflagged tweets

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
On a scale of "Very Unlikely" (1) to "Very Likely" (5), how likely are you to retweet or share the above tweet?	1.00	4.00	2.25	1.09	1.19	16
Field	Min	Max	Mean	Standard Deviation	Variance	Responses
On a scale of "Very Unlikely" (1) to "Very Likely" (5), how likely are you to retweet or share the above tweet?	1.00	5.00	2.44	0.86	0.75	16

Note. These results were taken from the responses of the control group.

These results indicated that the average numeric likelihood for users to share the flagged tweets averaged at about 1.225, but the unflagged tweets had a higher average of 2.345. To examine whether the presence of the context flag had an impact on the want to distribute specific tweets, we compared the responses from each tweet against their counterparts in the other group. For the pop culture tweet, there was a statistically significant difference found, $t(32) = -3.23$, $p = .002$. For the Samsung Mobile tweet, there was also a statistically significant difference found, $t(32) = -5.51$, $p < .0001$. Assuming the level of significance is 0.05, it is shown by the low p-values that the presence of the context flag had a significant impact on participants' likelihood to share the

two tweets. For the final question given to the treatment group asking whether their previous responses were affected by the addition of a context flag, approximately 83% of participants answered “yes”, as shown in Table 3. On the other hand, the final question given to the control group inquiring whether their previous responses would be affected by the addition of a context flag reported that around 69% of participants answered “yes”, shown in Table 4.

Table 3. Attitudes towards the impact of the presence of context flags on previous responses

#	Answer	%	Count
1	Yes	83.33%	15
2	Maybe	11.11%	2
3	No	5.56%	1
	Total	100%	18

Note. These results were taken from the responses of the treatment group.

Table 4. Attitudes towards the impact of the presence of added context flags on previous responses

#	Answer	%	Count
1	Yes	68.75%	11
2	Maybe	31.25%	5
3	No	0.00%	0
	Total	100%	16

Note. These results were taken from the responses of the control group.

Conclusion:

Based on the research conducted, it is indicated that there is a relationship between the presence of context flags on tweets and the distribution of misinformation on Twitter. Since it is shown from the survey results that there is statistical significance between the calculated p-values, them being less than 0.05, it can be noted that the results have a low chance of being

brought on by chance. From this, it can be suggested that the presence of a context flag on social media platforms like Twitter may have the possibility to influence users' choices whether to share certain content online. In addition to those conclusions, since both groups had a majority of participants believe that their responses could be impacted by the presence of a context flag, it can be assumed that the presence of a flag does have the ability to reduce the spread of misinformation on Twitter. However, it is important to note that this experiment has limitations due to the use of an online survey through only one platform, which may not equate to users' real behavior.

References

Allen, J. N. L., Martel, C., & Rand, D. G. (2021, December 2). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. https://doi.org/10.31234/osf.io/57e3q

Coleman, K. (2023). Introducing Birdwatch, a community-based approach to misinformation. Twitter. Retrieved May 3, 2023, from https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

Colliander, J. (2019). "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. Computers in Human Behavior, 97, 202–215. https://doi.org/10.1016/j.chb.2019.03.032

Lanius, C., Weber, R., & MacKenzie, W. I. (2021). Use of BOT and content flags to limit the spread of misinformation among social networks: A behavior and attitude survey. Social Network Analysis and Mining, 11(1). https://doi.org/10.1007/s13278-021-00739-x

@rapalert10. (2023, April 18). 🚨Ariana Grande teams up with Nicki Minaj and ice spice for the 'princess diana' alternate remix. https://t.co/cv7mh1rdck. Twitter. Retrieved May 3, 2023, from https://twitter.com/rapalert1o/status/1648353185608904704?s=42&t=C-pFGNYcQ3o9JmOfyToicg

@SamsungMobile. (2023, April 14). There's no dark side of the Moon with the #GALAXYS23 ultra. capture your night skies and share with us by replying to this thread with #sharetheepic. Twitter. Retrieved May 3, 2023, from

https://twitter.com/samsungmobile/status/1646875453473583105?s=42&t=C-pFGNYcQ3o9JmOfyToicg

Wright, A., & Upcraft, S. (2023). Qualtrics Research Survey. msu.co1.qualtrics.com. Retrieved May 4, 2023, from https://msu.co1.qualtrics.com/jfe/form/SV_afuWvCFf1MgPaJg