A Comparison of Account-Focused and Content-Focused Warnings on User Trust of Twitter Content

Ronald Thompson†, Santana Koring'ura†, Marshini Chetty*, and Daniel Votipka†
† *Tufts University, *University of Chicago*
{ronald.thompson,santana.koring_ura,daniel.votipka}@tufts.edu
marshini@uchicago.edu

Abstract

To date, social media platforms have employed warnings about posted content to combat disinformation campaigns that attempt to strategically mislead users for political gain. In this paper, we consider account-focused warnings to existing content-focused approaches to slowing a disinformation campaign's spread. We reviewed 65 anti-disinformation resources to identify possible account-focused warning features. Then, we created mock warnings for Twitter using these features and surveyed 942 users on Prolific. We found account-focused warnings are more effective at reducing perceived content accuracy. Participants were also generally comfortable with both account-focused and content-focused anti-disinformation warnings. However, perceived content accuracy and concerns were affected by partisanship. Based on our study, we recommend account-focused warnings as a useful tool against disinformation campaigns on social media, if used responsibly and transparently.

1 Introduction

Political discourse has fundamentally changed with the advent of social media [24]; bots have become more common as a means to influence political conversations [25], and politicians that previously would not have received much media attention have been able to garner attention [10]. At the same time, misleading and false information has proliferated on these platforms at a scale challenging their ability to stay ahead of bad actors attempting to sway discourse [16]. One particularly challenging problem is disinformation campaigns, sometimes referred to as coordinated inauthentic behaviour [8]. In these campaigns, multiple accounts run by one or more users purposefully spread false or misleading information in a coordinated fashion intended to maximize their efforts' reach and effect. [2]. This differs from misinformation as the primary goal is to sow confusion and discord, while accounts spreading misinformation may not do so with malicious intent.

Efforts to disrupt disinformation campaigns from spreading on social media have been focused mainly on content moderation. In mid-2017 Facebook [15] and Twitter [7] started taking a more proactive approach and began to add warnings to some types of content. Now content warning labels are common on both platforms and used in a variety of contexts.

Social media platforms have taken actions directly on accounts engaged in disinformation campaigns using methods such as temporary bans or de-platforming. However, these actions and the reasoning behind them are not immediately visible to users, but instead are published in lengthy periodical reports on corporate sites [14, 23]. Prior work investigating users' methods for determining message credibility suggests presenting this information directly to users in a warning may be beneficial [1, 3, 5, 5, 9, 11, 21]. In this paper we seek to evaluate the possible benefit of account-focused warnings on social media platforms. We test the following hypotheses:

H1: Account-focused warnings are more effective than content-focused warnings at reducing user trust in associated content.

H2: Users are more comfortable with a social media platform's use of account-focused warnings than content-focused warnings.

Hypothesis *H1* directly compares account-focused warnings to traditional content-focused ones. To evaluate user trust in associated content, we consider whether the warning affects the users' belief that the content is accurate and the users' reported likelihood to share the content. By inducing a higher level of uncertainty about the content on initial review, an effective warning helps support more critical user engagement with the content and can slow disinformation campaigns.

We also seek to understand whether users are comfortable with account-focused warnings (*H2*) and whether there are alternative warning presentations which might increase comfort. If users are uncomfortable with these warnings, they are less likely to trust the platforms' assessment of the account or content, and therefore less likely to believe the information included in the warnings [3, 11].

We first conducted a review of 65 existing anti-disinformation resources to establish a set of possible account-

and content-focused warning features for comparison. Next, we surveyed 942 US social media users to measure account- and content-focused Twitter warnings' effect on user perceptions of message trustworthiness. We measured perceived trustworthiness by asking about the content's accuracy and by asking directly about the warning's effect.

Our results show that warnings highlighting accounts which misrepresent themselves or regularly engage in bot-like or coordinated questionable behaviors were more likely to cause participants to question the accuracy of quoted content. Additionally, while content partisanship played a significant role—as might be expected—its impact on participants' responses and who their distrust was directed toward differed between Trump voting participants and Biden ones. Biden voting participants were more likely to distrust Republican quoted content, regardless of the warning. Whereas, Trump voting participants were more likely to report higher trustworthiness in quoted content regardless of the content's partisanship.

2 Study Design

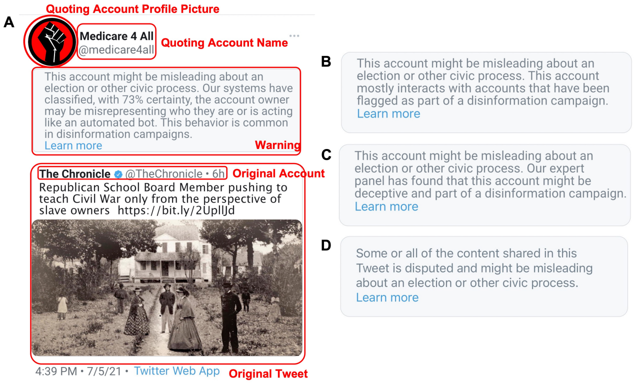
In order to answer these research questions we conducted a two part study. The first part was a comprehensive resource review of disinformation tools followed by an user survey that looked at the effectiveness of account warnings.

2.1 Resource Review

To inform the design of candidate warnings for *H1*, we looked to existing anti-disinformation resources that provide users with information about disinformation campaigns and their associated accounts. We identified a canonical list of 78 anti-disinformation resources compiled by the RAND Corporation [22]—an organization with deep disinformation expertise [4, 13]. We paired the RAND list to 65 resources for our analysis after accounting for dead links and broken resources.

Two researchers performed an iterative initial coding [6, 19] of all 65 resources. For each resource, the researchers identified the *Type of Information* given to a user (e.g., account details, how to spot disinformation, fact-checking) and its *Presentation Medium* (e.g., text, dashboards, videos, graphs). To assess inter-rater reliability, we used Krippendorff's Alpha (α) as it accounts for chance agreements [12]. This process was repeated for four rounds when sufficient reliability was achieved for all variables ($\alpha > 0.80$) [12].

We then created five higher level categorizations for *Type of Information* (bots/misrepresentation, account networks, investigations/discussions, sourcing/posts, and miscellaneous) and *Presentation Medium* (text, quantitative measure, changes to page, and content). We found additional categories that were interesting, but they were not as relevant to our investigation as they could not be easily presented in the form of a warning.



Medicare 4 All
@medicare4all

This account might be misleading about an election or other civic process. Our systems have classified, with 73% certainty, the account owner may be misrepresenting who they are or is acting like an automated bot. This behavior is common in disinformation campaigns.
Learn more

The Chronicle @TheChronicle • 6h
Republican School Board Member pushing to teach Civil War only from the perspective of slave owners https://bit.ly/2UplIJd

4:39 PM • 7/5/21 • Twitter Web App

This account might be misleading about an election or other civic process. This account mostly interacts with accounts that have been flagged as part of a disinformation campaign.
Learn more

This account might be misleading about an election or other civic process. Our expert panel has found that this account might be deceptive and part of a disinformation campaign.
Learn more

Some or all of the content shared in this Tweet is disputed and might be misleading about an election or other civic process.
Learn more

Figure 1: **A** Example of tweet and warning with *Type* as *Behaviour* and *Presentation* as *Quantitative*. **A** also shows the parts of the quoted tweet in red. **B-D** are example warnings with *Presentation* as *Text*. **B** has *Type* as *Connections*, **C** as *Qualitative*, and **D** shows Twitter's content-focused warnings.

2.2 User Study

Next, we ran a user survey where participants were shown a series of tweeted stories with various warnings applied depending on the treatment condition. We focused specifically on Twitter as it is a well known platform that is actively trying to address disinformation campaigns. In this section, we describe the design of the story presented to participants (See Figure 1 for an example), along with the survey design.

Content & Account. All messages were shown in the form of a quote tweet, where a story is originally shared by a mock news organization named to avoid partisan signals. Twitter's Quote Tweet functionality is used as part of disinformation campaigns since it amplifies a divisive message. In addition to this baseline structure, we varied the quoting account name and profile picture, as well as the original content. Specifically, we tested three account/content variations to test the effect of tweet partisanship: Left-Wing Account/Anti-Republican Content (*Left-Wing*), Right-Wing Account/Anti-Democrat Content (*Right-Wing*), or Science Account/Public Health Content (*Public Health*). In the partisan treatments, we selected content portraying the other partisan group negatively, as negative messages are likely to be more divisive—a common goal in disinformation campaigns [20]. For example, in Figure 1.A, we present a *Left-Wing* tweet. A left-wing account name (Medicare4All) and picture (a BlackLivesMatter logo) are shown tweeting a negative story about Republicans. For each of type, we also included two possible stories, one pulled from the headlines and fact checked, though not widely shared or covered nationally, and the other made up by the researchers. We chose to vary content veracity to ensure any effect on perceived trust was not caused by our choice of mock tweets. We used a full-factorial combination of content veracity and partisanship to create six mock quote tweets. Each participant was shown one tweet from each partisan group (i.e., Left-Wing, Right-Wing, and Public health), ordered randomly.

Warnings. For our mock warnings, we began with the text of Twitter's existing content-focused warning (Figure 1.D) and modified based on our results from Section 2.1 according to two variables: *Type* and *Quantitative Included*. For *Type*, we tested four options: 1) *Account Behaviour & Misrepresentation* (Figure 1.A), actions taken by the account exhibit bot-like behaviors or misrepresentation to influence target audience (e.g., frequently changing account name to affiliate with opposing political groups); 2) *Account Connections* (Figure 1.B), associated with other accounts identified as part of a disinformation campaign; 3) *Qualitative Evaluation* (Figure 1.C), deemed by an expert panel to be part of a disinformation campaign; 4) *Twitter's content warning* (Figure 1.D), the participant was shown Twitter's current content warning. The *Quantitative Included* variable was binary, with the warning either including a quantitative measure (i.e., a numerical measure of the warning's accuracy) or not. Therefore, for each of our four types, we created two possible warnings, giving us eight warning conditions. We also included a control treatment in which participants were not shown any warning. The full set of conditions can be found in our supplemental material.[1]

Survey Procedure. Participants were shown three tweets, one tweet from each of the three partisan content types (i.e., Left-Wing, Right-Wing, and Public Health). Each tweet was randomly assigned one of nine warning conditions. We ensured the conditions were balanced and the tweet order was randomized to avoid ordering effects [18]. All the warning/partisan content combinations were shown to at least 51 participants.

After each tweet, participants were asked to rate the tweet's accuracy and their likelihood to share it on a five-point Likert scales. These questions were chosen to evaluate hypothesis *H1*. After seeing all three tweets, we showed participants the tweets again in the same order and asked which parts of the tweet affected their perception of the content s accuracy. If the participants was shown a warning, we also asked them to rate their perception of the warning's effect on a four-point Likert scale. When asking about the Public Health tweet, we included an attention check asking the participant to select the name of the news organization (which was visible on-screen) that reported the quoted content.

We also asked participants to consider their comfort with account- and content-focused warnings generally (*H2*). Participants rated on a five-point Likert-scale scale from "Not at all comfortable" to "Very comfortable", their comfort with content- and account-focused warnings and explained their reasoning in a free-response question. Prior to showing these questions, we defined account- and content-focused warnings and showed examples of each. We randomized account- and content-focused question ordering [18].

After completing the survey, participants were given links to disinformation education resources and told which headlines shown during the survey were factually correct.

Recruitment. We conducted our survey on Prolific [17]. To avoid self-selection biases, we did not mention disinformation campaigns explicitly in our study description. We limited participation to users over 18 years old and located in the US. To keep vote choice balanced, we recruited in multiple rounds and filtered based on vote choice. Participants were paid $3 upon completion (approximately $12/hour).

Ethics. Our study was approved by Tufts' and the University of Chicago's Institutional Review Boards. All participants provided informed consent at the start of the survey. We did not collect identifiable information beyond participant Prolific IDs, which were only used for participant compensation.

3 Results

In total, 942 participants completed our survey, passed the attention check, and did not provide nonsensical or unresponsive answers to free-text questions. For brevity, we present the most salient results from our analysis of these 942 responses.

3.1 Account-focused vs. Content-focused (H1)

To evaluate *H1*, we first compare responses between participants shown account- and content-focused warnings. For brevity, we will only discuss perceived accuracy (regression results give in Table 1), as participants' reported likelihood to share content closely matched perceived accuracy.

Account-focused warnings lower perceived accuracy. Account-focused warnings had a significant effect on participants' perception of the quoted tweet's accuracy. Participants were less likely to perceive a quoted tweet as accurate if an account-focused warning was shown instead of a content-focused warning (OR=0.78, p=0.016). Specifically, warnings of *Type Behaviour* had the largest effect on accuracy (OR=0.68, p=0.002). Account-focused warnings' effects were also evident in our qualitative responses. 28.13% of participants said the warning was informative when an account-focused warning was shown, compared to 23% of participants shown content-focused warnings.

Responses vary by party. While *Behaviour* warnings were effective in decreasing perceived accuracy, we also observed significant partisan effects. Overall, partisan quoted tweets (i.e., *Left-Wing* or *Right-Wing*) were less likely to be viewed as accurate than *Public Health* tweets (OR=0.59, p<0.001 and OR=0.32, p=0.59, respectively). Biden voters rated 68% of *Right-Wing* content as inaccurate, but only 41% said *Left-Wing* content was inaccurate. In a majority of cases (62%), Biden voters explained that prior knowledge and the warning confirmed their pre-existing beliefs. For instance, P58 said, "Honestly, I'm more likely to believe that a warning has

[1] Supplemental material can be found at https://osf.io/v4ndw/?view_only=8077fc858a1a448f9df59d2ade1287c5.

	Variable	Value	OR	p-value	CI
Warning	Presentation	Text	-	-	-
		Quantitative	0.85	0.058	[0.74, 0.98]
	Type	Twitter	-	-	-
		No Warning	**1.68**	0.001	[**1.3**, **2.18**]
		Behaviour	**0.68**	0.002	[**0.56**, **0.83**]
		Connections	0.86	0.223	[0.71, 1.05]
		Qualitative	0.82	0.1	[0.67, 1.0]
Post	Partisanship	Public Health	-	-	-
		Left-Wing	**0.59**	**0.0**	[**0.49**, **0.72**]
		Right-Wing	**0.32**	**0.0**	[**0.26**, **.39**]
	Veracity	False	-	-	-
		True	**1.43**	**0.0**	[**1.23**, **1.66**]
Participant	2020 Vote	Biden	-	-	-
		No Response	**1.67**	**0.026**	[**1.14**, **2.45**]
		Other	**1.35**	**0.032**	[**1.07**, **1.7**]
		Trump	**1.33**	**0.016**	[**1.1**, **1.62**]
	Partisanship of News	Bi-Partisan	-	-	-
		Moderate	**0.62**	**0.013**	[**0.46**, **0.85**]
		Conservative	1.03	0.877	[0.74, 1.43]
		Liberal	0.93	0.663	[0.71, 1.22]
		None	0.9	0.477	[0.71, 1.14]
		Mixed	0.83	0.184	[0.65, 1.05]
	Social Media News Consumed	Low	-	-	-
		High	**1.25**	**0.018**	[**1.07**, **1.46**]

* Significant effect – Base case (Odds ratio defined as 1)

Table 1: Results of perceived accuracy regression split by *presentation* and *type*. OR above one implies more likely to believe quoted content was accurate than the base case.

more validity on right leaning accounts/sources." Conversely, Trump voters, focused their distrust toward the warnings themselves with 44.88% saying the warnings were biased or inaccurate. Trump voters were more likely to view the quoted content as accurate than Biden voters (OR=1.33, p=.016).

3.2 Comfort with warnings (H2)

Participants reported being *Somewhat* or *Very Comfortable* with both account-focused (56%) and content-focused (63%) warnings. Many participants (72%) pointed to the usefulness of both warning types when asked why they were or were not comfortable. This was the dominant response among participants and across parties. Participants reasons for believing the warnings were useful varied from helping figure out what news is accurate ("it allow us to know which news is true or false" P645) or protecting them from malicious actors ("If someone following me seems to be a scam, or bot account I would like to know because I do not feel comfortable with them seeing my information." P659) to improving the social media information ecosystem by holding malicious actors accountable ("If an account continuously spreads false information then they should be held accountable…" P355). These responses were similar in nature to participants' reasons for finding specific warnings we tested useful in Section 3.1. Many participants had concerns with warnings even if they said they were comfortable with them and that warnings might be helpful. We found the following three major concerns.

Distrust of social media companies. A large portion of participants distrusted social media companies themselves (41%) believing these companies were biased against specific populations. P600 explained, "I do not trust the social media companies to be arbiters of truth. I believe they are agenda driven and are passing their opinions off as facts."

(In)Accuracy. Participants were also concerned about warning inaccuracy (16%). Some participants personally being flagged, or knowing people who had been, with warnings they thought were inaccurate. P494 said, "Automation has flagged myself and my friends for tons of innocuous content, particularly satire. There is no room for nuance with AI." Participants also raised accuracy concerns around the inherent changing nature of information over time, particularly in science. As P129 explained, "Science is full of disagreements. Who's to say which expert is right or wrong?… Hell, Einstein said Quantum Physicists were wrong."

Clarification and more information. Finally, participants mentioned that more information was needed (4%) to clarify how warnings are being applied and how to use this information. P617 summed this concern up well, "If something is inaccurate, you need to tell me what's inaccurate. 'Some or all of the content conflicts with guidance from public health experts' is meaningless. Eating pancakes with syrup 'conflicts with guidance from public health experts."'

Political lean affects warning comfort. We also observed that these concerns about warnings were correlated with participants' partisan lean. While a plurality of Trump voting participants said warnings were useful (47%), the vast majority of Biden voters reported warnings as useful (87%). Trump voters also reported concerns more often than Biden voters with abuse (39% vs. 13%) and accuracy (18% vs. 15%). We also observed that participants who reported consuming only conservative sources (e.g., New York Post, Fox News, and Sean Hannity Show) were less likely to be comfortable with both account-focused (OR: 0.36, p< 0.001) and content-focused (OR: 0.43, p: 0.002) warnings.

4 Conclusion

Account-focused warnings appear to be a small step in the right direction. These types of warnings would be a helpful tool for social media companies as they try to slow possible disinformation campaigns, while taking time to validate and verify an account's participation in a disinformation campaign. However, given distinguishing dis- and mis-informaiton is incredibly hard and many of these issues are highly partisan in nature, these warnings are not a silver bullet and should be considered in the context of broader systemic changes.

References

[1] Mikhail Anufriev, Cars Hommes, and Tomasz Makarewicz. Simple forecasting heuristics that make us smart: Evidence from different market experiments. 17(5):1538–1584.

[2] Yochai Benkler, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*.

[3] David K Berlo, James B Lemert, and Robert J Mertz. Dimensions for evaluating the acceptability of message sources. *Public opinion quarterly*, 33(4):563–576, 1969.

[4] Elizabeth Bodine-Baron, Todd Helmus, Andrew Radin, and Elina Treyger. *Countering Russian Social Media Influence*. RAND Corporation, 2018.

[5] Shelly Chaiken. The heuristic model of persuasion. In *Social influence: The Ontario symposium, Vol. 5.*, Ontario symposium on personality and social psychology, pages 3–39. Lawrence Erlbaum Associates, Inc.

[6] Kathy Charmaz. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE. Google-Books-ID: v1qP1KbXz1AC.

[7] Colin Crowell. Our approach to bots and misinformation.

[8] Nathaniel Gleicher. Coordinated inauthentic behavior explained.

[9] Brian Hilligoss and Soo Young Rieh. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. 44(4):1467–1484.

[10] Sounman Hong, Haneul Choi, and Taek Kyu Kim. Why do politicians tweet? extremists, underdogs, and opposing parties as political tweeters. 11(3):305–323. _eprint: https://misclibrary.wiley.com/doi/pdf/10.1002/poi3.201.

[11] C.I. Hovland, I.L. Janis, and H.H. Kelley. *Communication and Persuasion: Psychological Studies of Opinion Change*. Greenwood Press, 1953.

[12] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. 30(1):61–70. Publisher: SAGE Publications Inc.

[13] Miriam Matthews, Katya Migacheva, and Ryan Andrew Brown. Superspreaders of Malign and Subversive Information on COVID-19: Russian and Chinese Efforts Targeting the United States. Technical report, RAND Corporation, April 2021.

[14] Meta. Coordinated Inauthentic Behavior Archives.

[15] Adam Mosseri. Working to stop misinformation and false news.

[16] 1615 L. St NW, Suite 800Washington, and DC 20036USA202-419-4300 {\textbar} Main202-857-8562 {\textbar} Fax202-419-4372 {\textbar} Media Inquiries. The future of truth and misinformation misc.

[17] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[18] Harry T Reis, Harry T Reis, Charles M Judd, et al. *Handbook of research methods in social and personality psychology*. Cambridge University Press, 2000.

[19] Johnny Saldana. *The Coding Manual for Qualitative Researchers*. SAGE. Google-Books-ID: V3tTG4jvgFkC.

[20] Kate Starbird. Disinformation's spread: bots, trolls and all of us. 571(7766):449–449. Bandiera_abtest: a Cg_type: World View Number: 7766 Publisher: Nature Publishing Group Subject_term: Society, Information technology.

[21] S. Sundar. The MAIN model : A heuristic approach to understanding technology effects on credibility.

[22] The RAND Corporation. Tools that fight disinformation misc. `https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html`.

[23] Twitter. Information Operations - Twitter Transparency Center.

[24] Rachelle Vessey. Zappavigna, m. (2012). discourse of twitter and social media: How we use language to create affiliation on the web. london: Bloomsbury. In Jesús Romero-Trillo, editor, *Yearbook of Corpus Linguistics and Pragmatics 2015: Current Approaches to Discourse and Translation Studies*, Yearbook of Corpus Linguistics and Pragmatics, pages 295–299. Springer International Publishing.

[25] Samuel C Woolley. Political communication, computational propaganda, and autonomous agents - introduction. page 9.