The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets

Orestis Papakyriakopoulos
Center for Information Technology Policy
Princeton University
orestis@princeton.edu

Ellen Goodman
Rutgers Law School
Rutgers University
ellgood@rutgers.edu

ABSTRACT

Social media platforms are performing "soft moderation" by attaching warning labels to misinformation to reduce dissemination of, and engagement with, such content. This study investigates the warning labels that Twitter placed on Donald Trump's false tweets about the 2020 US Presidential election. It specifically studies their relation to misinformation spread, and the magnitude and *nature* of user engagement. We categorize the warning labels by type – "veracity labels" calling out falsity and "contextual labels" providing more information. In addition, we categorize labels by their rebuttal strength and textual overlap (linguistic, topical) with the underlying tweet. We look at user interactions (liking, retweeting, quote tweeting, and replying), the content of user replies, and the type of user involved (partisanship and Twitter activity level) according to various standard metrics. Using appropriate statistical tools, we find that, overall, label placement did not change the propensity of users to share and engage with labeled content, but the falsity of content did. However, we show that the presence of textual overlap in labels did reduce user interactions, while stronger rebuttals reduced the toxicity in comments. We also find that users were more likely to discuss their positions on the underlying tweets in replies when the labels contained rebuttals. When false content was labeled, results show that liberals engaged more than conservatives. Labels also increased the engagement of more passive Twitter users. This case study has direct implications for the design of effective soft moderation and related policies.

CCS CONCEPTS

• **Human-centered computing → Social network analysis**; Empirical studies in HCI.

KEYWORDS

content moderation, misinformation, warning labels, political discourse, Trump

ACM Reference Format:
Orestis Papakyriakopoulos and Ellen Goodman. 2022. The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3485447.3512126

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512126

1 INTRODUCTION

Western democracies have experienced a sharp increase in the speed and quantity of false content spreading through their information networks [9]. The salience of false statements in public discourse has been described as "post-truth" [39] and "information disorder" [52], leading to an epistemic fracture where large portions of society diverge in their understanding of reality. False information can be categorized as disinformation (intentionally false) or misinformation (unintentionally false) [16, 17]. For ease of reference, we use the term misinformation to cover all conditions of falsity. Platforms are under pressure to demote the salience of misinformation through "soft moderation," or to remove it entirely from circulation through "hard moderation" [42]. One form of soft moderation is the placement of warning labels on content containing misinformation. This practice was adopted by several social media platforms, including Twitter, which in 2020 started labeling misinformation related to the COVID-19 pandemic [2, 3].

Problem Statement. Although platforms used soft moderation practices intensively during the 2020 US Presidential election [1, 45], little is known about the effectiveness of the practice or how to operationalize it for greatest impact. To generate additional knowledge, we study Twitter's decision to assign warning labels to some of former-President Donald Trump's false tweets from election day, 3 November 2020, until the day of his account's permanent ban on 8 January 2021. During these months, Trump claimed falsely that extensive election fraud had taken place and that this fraud turned the election results against him. Understanding the relationship of Twitter's interventions to Trump's false statements is especially important because months after the elections, polling showed that one-third of Americans believed Trump's claims to be true [12], providing evidence of sustained epistemic fracture. Preliminary evidence from research studies [42, 57] shows that the use of warning labels on Trump's false tweets statistically coincides with an increase in the dissemination of, and user engagement with, the underlying content. Building on these studies, we analyze in greater detail the relationship between the deployed warning labels, misinformation spread, and the magnitude and type of user engagement. We find a more nuanced story and are able to identify differences among different kinds of labels, users, and types of engagement. We answer the following question:

RQ: How did different warning labels placed on Donald Trump's false tweets impact (a) misinformation spread, (b) magnitude, and (c) type of user engagement?

Original Contributions.

- We collect 1241 Tweets from Donald Trump's Twitter account between 30 October 2020 and 8 January 2021, their engagement metrics, and 2,399,366 corresponding user replies. We classify tweets as soft-moderated or not, and as containing three types of misinformation: fraud related, election victory related, and ballots related. We categorize warning labels by type (veracity or contextual), rebuttal strength, and their linguistic and topical overlap with the associated tweets.
- We find that, overall, when controlling for misinformation, warning labels did not result in a statistically significant increase or decrease in misinformation spread and user engagement.
- By using robust linear regression analysis, we find that veracity labels were related to more user engagement and sharing compared to contextual labels. We also find that labels with linguistic or topical overlap were associated with a decrease in users liking, retweeting, quote tweeting, and replying to labeled content.
- We study *who* was mobilized by the different labels and *how*. We categorize 71,338 random users by partisanship following the work of Barbera [8]. Since the median user on Twitter is primarily passive [54], and rarely creates content on the platform, we also calculate users' tweeting activity in our dataset. We then compare the distribution of users replying to moderated and unmoderated tweets containing misinformation, and conclude that tweet moderation was associated with a mobilization of less active users. Furthermore we find that labeled tweets were more likely to mobilize liberal than conservative users.
- By applying a structural topic modeling algorithm [40] to the tweet replies, and quantifying their toxicity level using Jingsaw's Perspective API [5], we locate divergences between the nature of discussions under soft-moderated and unmoderated tweets. We find that users employed more toxic language when commenting on labeled tweets. We also find that the presence of rebuttals in labels was associated with more discussion about the underlying claims in replies across the partisan spectrum - discussion that reflected partisan priors, but employed less toxic language.
- Based on the results, we argue that an optimal use of warning labels, providing strong rebuttals and textual overlap between label and tweets, can be a plausible way to mitigate misinformation spread and toxic user engagement. Given this, we discuss the policy and design implications of the study for content moderation and highlight pathways for future research.

2 BACKGROUND & RELATED WORK

2.1 Fact-checking warnings and content labels are of disputed efficacy

On social media, soft moderation practices encompass every action taken that informs users about problematic content, limits its visibility, or imposes specific limitations on users in engaging with it [57]. Among of these practices are fact-checking warnings and/or content labels, which are often used as de-biasing tools to get people to relinquish their beliefs and cease propagating false information. Labels provide a platform's or a third party's assessment of the truth or falsity of a statement or additional information to contextualize or clarify the content. The third party assessors may be news organizations, fact-checking organizations, or, in the case of election information, election officials. From a user interface perspective, labels are *"any kind of partial or full overlay on a piece of content that is applied by platforms to communicate information credibility to users"* [4]. Content labels can fall into two categories: 1. *veracity*, which provide explicit information about credibility, including a corrective statement, and 2. *contextual labels*, which provide more information on a topic.

While some studies have found that using labels works to reduce misinformation [14, 15, 19, 30, 55], others have found no effect at all [22], and still others have identified boomerang or "backfire" effects [31, 33]. The divergences in the effectiveness of content labels may be due to a number of psychological contributions. For example, false information tends to be "stickier" when it conforms to preexisting beliefs that are highly charged or salient to a person's worldview. In such cases, *"a correction can be interpreted as an attack on one's core beliefs or tribal identity and thus be ineffective or backfire"* [35]. Once they have adopted false beliefs, people find it difficult to change them if the falsehoods are tied up with personal and group identity [13, 49], with cognitive mechanisms such as *conformation bias* [58], *belief persistence* [48], or *story coherence bias* [27, 53] being activated.

By the most optimistic measures, a correction will at most halve the references to misinformation by those who have been exposed to it [53], with researchers finding that accuracy prompts [15] or prompts that exploit specific cognitive mechanisms can be successful [30]. However, the average effect of a correction does not seem to be high. One meta-analysis of debiasing studies by Chan et al. [11] found that debunking succeeded in neutralizing the effects of misinformation only 11.78% to 19.45% of the time. This analysis also found that debunking was less effective when people had already generated explanations supporting the false information than when they had not and that debunking was more effective when it provided new information enabling the update of previously-generated explanations. A meta-analysis of fact-check studies showed "mixed" evidence for efficacy in changing user beliefs [51]. Walter et al. examined 30 studies on fact-checking and found that it *"positively affect[s] beliefs, irrespective of political ideology"*. While there was significant heterogeneity in outcomes across political leanings, the effects overall were characterized as "quite weak" and they "gradually [became] negligible the more the study design resemble[d] a real-world scenario of exposure to fact-checking." Kaiser et al. [23] investigated the effects of contextual and interstitial warnings, but found that either users completely ignored the labels, or there weren't any significant behavioral changes related to their appearance. Studies of labels specifically in the social media context have not found labels to be particularly effective in changing beliefs or reducing engagement with misinformation [37]. A qualitative study of a small sample of users revealed significant distrust of, and irritation with, labels warning of misinformation [41]. An experiment comparing perceptions of COVID misinformation tweets that had labels [44] with those that had "warning covers" obscuring the underlying content found that the labels were ineffective, but the covers were effective [44]. There is some evidence that content labels can have unintended "backfire effects" [33, 47]. Corrections can paradoxically lead to increased belief in the underlying false information [35]. People may resist the influence of others and seek to reassert their freedom by rejecting the correction, in a phenomenon known as psychological "reactance" [46]. Corrections are less likely to be effective if they remind people of social difference or out-group status [32].

The efficacy of warning labels on social media is also influenced by additional features that are platform- and group- specific. For example, engagement with misinformation that is captured in metrics such as likes and retweets sends signals to users that others have validated the misinformation; these validations in themselves make the misinformation more powerful [7]. Perceived consensus can also lead to an illusory truth effect. Lewandofski et al. [26] showed that people are more likely to believe propositions that others believe, while a person holding an unpopular opinion is more likely to commit to it if others share it [21]. Especially when people share views, *"even when they all have the same information—these views tend to be held with more confidence"* [43]. It is evident therefore that whether and how labels should be used to reduce misinformation spread and user engagement remains a subject of debate. The detailed analysis of Twitters' labels on Donald Trump's false tweets can provide valuable information in this inquiry.

2.2 The chronology of Trump's account moderation

In October 2020, Twitter announced that it would place warning labels on tweets of public figures with more than 100,000 followers that contained specific types of misinformation or problematic speech [1]. These categories included false statements about the outcome of the elections, as well as calls to violently or non-violently interfere with the election process. On the night of 3 November, election day, Donald Trump violated these guidelines by claiming that there were fraudulent attempts to manipulate election results, a statement that official authorities disputed. Thereafter, he generated hundreds of tweets in which he either claimed he was the winner of the electoral process, or that there had been election fraud related to polling, vote counting, or mail-in ballot collection. Twitter soft-moderated many of these tweets, assigning various types of warning labels on most (although not all) of Trump's tweets that violated the guidelines. Donald Trump's account was finally removed by the platform on 8 January 2021, after his supporters attacked the US Capitol.

Earlier studies found that the labeling of Trump's tweets was associated with an increase in the spread of misinformation and user engagement. Zanettou [57] studied a set of labeled and non-labeled tweets on the platform, including Trump's, and found that labeled tweets had more user engagement compared to tweets without warning labels. He also found that most replies to tweets with warning labels were debunking false claims, mocking the tweet's creator or further supporting false claims. Similarly, Sanderson et al. [42], who studied Trump's tweets exclusively, found that Twitter's hard moderation of blocking engagement with tweets limited their spread on the platform, soft moderation that only flagged tweets with a warning label actually increased dissemination. However, they also found that messages blocked from engagement on Twitter were posted more often and became more popular on other online platforms than messages that were just labeled by Twitter or that received no intervention at all. These studies provide a preliminary understanding of the relation of misinformation spread, user engagement and warning labels. This study, taking into consideration these results, seeks to uncover how specific properties of the warning labels were related to specific patterns of misinformation dissemination and user engagement. In this way, it seeks to generate useful knowledge on whether and how to design warning labels in the future.

3 DATA & METHODS

3.1 Data Collection

We systematically crawled Trump's Twitter account from 1 November until the day of its removal. We collected the number of quote tweets, retweets and likes that each tweet generated, as well as the textual information of the warning labels. Overall, we collected 1241 tweets, on which Twitter had placed 10 unique warning labels (Appendix, Figure 4). Having obtained the ids of Trump's tweets, we used the full-archive endpoint of Twitter's researcher API, and collected the reply count to each of these tweets. We also collected a stratified sample of the actual replies, creating a corpus of 2,399,366 replying tweets out of the 35,556,679 generated during the specific period. From the collected tweets, we focused on the 220 and 623 single-labeled and unlabeled original tweets of Trump (excluding his retweets and quote tweets). There were 15 tweets bearing more than one label for which user interactions were deactivated, which we did not take into consideration in our analysis.

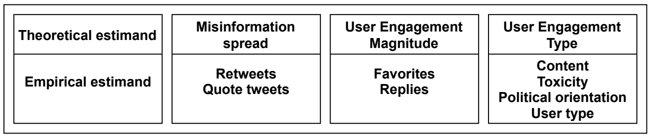
As a next step, we characterized the various labels placed on Trump's tweets. We labeled them as *veracity labels*, if they called out falsity, or as *contextual*, if they provided more information. We further generated three variables to describe the labels. The first was rebuttal strength to quantify the magnitude of correction a label provided. A label could have no rebuttal strength (0), moderate rebuttal strength (1) or high rebuttal strength (2). A label would have no rebuttal strength if e.g. the label was actually unrelated to the content of the tweet. It would have moderate rebuttal strength if e.g. Trump tweeted that his party won the election in Pennsylvania, while the label said that Biden won the election overall. It would have high rebuttal strength if e.g. Trump claimed that he won the election, while the label said that Biden won.

The second variable was *linguistic overlap* to signify that the label used the exact same vocabulary as did the tweet. For example, the variable would have the value one when the word "fraud" appeared both in the label and the tweet. The third variable was *topical overlap* to signify that the label and tweet referred to the same issue, but with different wording. For example, the variable would have the value one when the word "steal" appeared in Trump's tweet, while the label referred to election "fraud". Table 3 (Appendix) provides examples of the deployed classification scheme. Two coders manually labeled the corpus of flagged tweets, reaching satisfactory agreement (Krippendorf's $\alpha = 0.81$). In cases of disagreement, coders reviewed labels together and came to an agreement for the final classification. Overall we detected 217 original single-labeled tweets with a veracity label and 7 tweets with a contextual label. For 79 tweets there was a linguistic overlap between label and tweet, in 126 cases there was a topical overlap and in 18 cases, there was no overlap at all. Furthermore, in 181 cases labels had a moderate rebuttal strength, in 18 a strong one, and in 19 no rebuttal strength at all.

To create the conditions for a quasi-experimental analysis, we manually labeled all Trump's tweets in terms of their content. Specifically, we identified the tweets containing misinformation, and categorized them in three categories: Whether he claimed there was election fraud, whether he claimed he won the election, and whether he referred to the dangers of mail-in ballots. Overall, we found 285 original tweets that contained misinformation, of which 85 were not labeled as such. The intercoder agreement for this step was $\alpha_k = 0.74$, with the coders reaching a consensus for classification differences. Since Twitter labeling was independent for each tweet and given the existence of a reach pool of variables describing the content of the tweets, we can deploy appropriate tools to investigate which properties were associated with misinformation spread and user engagement. It is important to note that the quasi-experimental design that follows cannot account for all potentially unobserved confounding variables [28, 38]. Therefore, the results of the study are mainly associative. Nonetheless, they provide information that has direct policy implications. The generated knowledge can be used in future research studies for designing appropriate randomized experiments and verifying causal associations. Since moderating Trump's account was a unique case study, associations should be tested in other social media contexts as well to be generalizable.

3.2 Defining estimands

Lundberg et al. [29] argue that the understanding of the theoretical targeted quantities studied (estimands) and their connection to empirical ones is a key part for transforming statistical evidence to theoretical knowledge.



Theoretical estimand	Misinformation spread	User Engagement Magnitude	User Engagement Type
Empirical estimand	Retweets Quote tweets	Favorites Replies	Content Toxicity Political orientation User type

Figure 1: Theoretical and empirical estimands in the study.

Next, we connect these theoretical estimands to empirical ones (figure 1). These are variables that we can measure, and we can study their relation to other variables. We claim that the number of retweets and quote tweets can depict how much users actually share misinformation, as these engagement mechanisms directly redistribute content through the network. We quantify the magnitude of user engagement with misinformation by the number of likes and replies each tweet received.

To understand the type of user engagement, we use and study four distinct empirical estimands: content, toxicity, political orientation, and user activity type. These estimands allow us to evaluate the relationship between warning labels and some qualitative features of political discourse. Content moderation is per se a political issue [20], and its effects should be understood in terms of how they affect political discourse. Content and toxicity as estimands allow us to examine how labels might influence the content and anger of political discussions, while user political orientation and activity level help to illuminate how different parts of the public respond to labeled content.

We quantify the content of replies by structural topic modeling. The content of user replies can show how warning labels might inform discussions about misinformation, and uncover themes in user engagement. Second, we measure the amount of toxicity in the user replies by using the Jigsaw perspective API. We investigate in this way whether discussions became more toxic when fact-checking took place, as findings of other studies have already shown [24, 57]. Third, we calculate the political orientation of users engaging in the discussions by using Barbera's generative model [8], which classifies users on the left-right scale based on their Twitter following network and has been updated for the 2020 elections. We can quantify how different labels might have mobilized partisans, thereby showing *who* was influenced the most by soft moderation and with what effect. Fourth, since prior research studies [34, 50] have shown that users with different activity levels externalize different behaviors, we count the number of replies each user generated, and use it to understand whether labels might activate specific types of users (passive, active) to engage in discussions. Prior research has shown [54] that the majority of users remain passive on the platform (passive users), rarely interacting with or generating content. In contrast, a small portion of users, 10% or less (active users), is responsible for generating the majority of content. This is important to take into consideration, since any associations found between warning labels and user behavior might not be generalizable to the total user population.

3.3 Methods

We investigate how the values of the empirical estimands changed given the prevalence and properties of warning labels by creating regression models and performing corresponding hypothesis testing.

To associate the change of the empirical estimands related to misinformation spread and user engagement with soft moderation practices, we calculate the median value of retweets, quote tweets, likes, and replies that users generated on labeled and unlabeled tweets. We perform pairwise Mann-Whitney U - tests (two-tailed) to locate statistically significant differences. To quantify associations with properties of the warning labels, we develop separate robust linear regression models that have as dependent variables the logarithmized number of retweets, quote tweets, likes, replies. We introduce the log transformation of the dependent variables to reduce the skewness in their distribution. We use robust linear regression over normal least squares estimation, since our data contain outliers in terms of engagement metric values. Robust linear regression uses an M-estimator that weighs in the contribution of each observation [56], hence being more resilient to such issues.To deal with issues of detected heteroscedasticity and autocorrelation, we use robust standard errors following the implementation of Andrews [6]. We elect as independent variables the existence of a label (veracity, contextual), the textual overlap between warning label

Table 1: Median interactions on Trump's tweets with and without warning labels. Starred results illustrate statistical significant difference between labeled and unlabeled tweets at $P < 0.05$, as calculated by the Mann-Whitney U test.

Median engagement	Nr. tweets	Likes	Retweets	Quote tweets	Replies
All					
Not Labeled	628	164,472	31,595	25,000	19,359
Labeled	224	**229,499***	**53,209***	**47,500***	**35,676***
Fraud					
Not Labeled	56	229,490	50,358	42,000	35,002
Labeled	169	245,831	54,881	48,000	36,064
Win					
Not Labeled	24	220,166	42,503	57,500	40,214
Labeled	56	246,402	55,257	58,500	40,920
Ballots					
Not Labeled	6	251,509	61,959	**67,500***	65,859
Labeled	13	189,842	42,808	43,000	27,068

and text (linguistic, topical overlap), and the label's rebuttal strength (no rebuttal, moderate rebuttal, strong rebuttal). In our models, we control for the date a tweet was created, since engagement metrics changed over time based on circumstances, and misinformation existing in Trump's tweets, regardless of being labeled or not. Since for specific levels of categorical variables we have limited amount of observations (e.g contextual labels, strong rebuttal strength), we expect that standard errors to be larger than usual. We still use as level of significance p=0.05, but focus on the effect size of independent variables when interpreting results rather than strict statistical hypothesis testing.

To quantitatively investigate the user engagement type we calculate the values of the corresponding empirical estimands: A. topics of tweets, B. toxicity in replies, C. user partisanship, D. user activity type. Because of computational restrictions and the limits imposed by the Twitter and Jigsaw APIs, we are not able to apply the topic modeling algorithm and the toxicity classifier on the total corpus of user replies, nor to calculate user partisanship for all users in our collected data. Instead, we create random samples of users and replies, based on which we generate our estimands. First, we calculate how many replies did each user in our corpus generate, calculating their activity level in the discussions. Next, we randomly sample 71,338 users, for which we calculate their partisanship score by applying Barbera's algorithm [8] on users' following network, as collected by the Twitter API. For running the structural topic model, we create a stratified sample of 200,000 user replies. We run the algorithm with spectral initiation, in order for it to automatically infer the number of topics in the corpus, yielding 64 distinct topics. We do not use further covariates in model optimization, since we prioritize using a larger corpus rather than modeling more complex associations in the data. We then use the trained model to predict the content of all tweets (284,725) created by the previously sampled 71,338 users. Furthermore, we use the Jigsaw API to calculate the probability of a comment being toxic for the above tweet corpus. In this way, we quantify all four empirical estimands related to user engagement type.

To uncover whether warning labels prevalence and type is associated with a change in the type of user engagement we proceed as follows. We generate the distributions of users who interacted with moderated and unmoderated tweets that contained misinformation both in terms of partisanship and activity level. By performing Anderson-Darling two-sample tests, we measure whether these distributions significantly deviate from each other for each concrete sub-case, i.e. across different label types, labels' textual overlap with the underlying tweets, and labels' rebuttal strength. We also calculate the percent of users existing beyond the interquartile range of the distributions, as a way to quantify how many extreme partisans created replies. Furthermore, we compare the number of replies belonging to each calculated topic for moderated and unmoderated tweets.

For topics related to misinformation, for which we find a substantial difference between labeled and unlabeled tweets in terms of user replies, we perform stance detection, in order to quantify how users supporting or disputing misinformation behaved across the different moderation conditions. For this, two coders manually label 266 user replies belonging to these topics as "supporting misinformation", "disputing misinformation", or "other" ($\alpha_k = 0.86$) and fine-tune further Kawintiranon al. [25] NLP architecture. The specific language Bert-based model was already fine-tuned on an election related corpus with the explicit purpose to be used for stance detection tasks related to the US 2020 Elections. Our final classifier has a satisfactory test-set performance, as it is illustrated by the test-set F-1 scores of 0.83 for "other" content, 0.78 for "supporting misinformation" replies, and 0.85 for "disputing misinformation." We classify then all replies belonging to these topics, and create a multinomial regression classifier that predicts users' propensity to argue in one of the three predefined ways. We include as independent variables the label type, label properties, the content of Trump's tweets, date and the user activity level and political orientation. Besides modeling the propensity of users to reply in a specific way, we calculate which factors are associated with generating more toxic replies. For this, we create a beta regression model that gives the probability of a reply being toxic given label type, label properties, date, user orientation and activity level.

4 RESULTS

4.1 Overview of differences between labeled and unlabeled tweets.

The first result of the study initially aligns to prior research findings [42, 57]: As Table 1shows, when comparing the sum of labeled and unlabeled tweets, the labeled ones were associated with greater user engagement in terms of likes and replies, and also greater sharing in terms of quote tweets and retweets.The overall difference in interactions was substantial, with labeled tweets being liked approximately 36% more, retweeted 70% more, quote tweeted 88% more, and the median number of replies labeled tweets generated was 84% higher than that of unlabeled tweets. The Mann-Whitney U tests also yielded that these differences were statistically significant at a $P < 0.05$ threshold. However, when controlling for the content of the tweets, these differences disappear. Comparing tweets that explicitly contained misinformation narratives, the statistical analysis revealed that users interacted with the labeled and unlabeled tweets at similar levels. For labeled and unlabeled tweets in which Trump claimed the existence of electoral fraud, labeled tweets were slightly more liked, retweeted, quote tweeted, and replied to, but none of these differences was statistically significant. For tweets in which Trump claimed he won the elections, labeled tweets generated a higher number of user interactions, albeit no difference was statistically significant. For ballot-related tweets, unlabeled tweets were actually associated with more user engagement and sharing, with the difference for quote tweets being statistically significant. These results illustrate that warning labels did not necessarily lead to an increase in misinformation spread and user engagement. Overall, they did not change the number of user interactions with the tweets. However, we found more complex relationships between the labels and user behavior, as we illustrate next.

4.2 Relation of labels, misinformation spread & magnitude of engagement

The robust linear regression model results provide a more detailed picture about the relationship between warning labels and user interactions in terms of likes, replies, retweets and quote tweets (Figure 2).

Regardless of interaction type these associations followed a uniform pattern (Appendix, Table 4). In all cases, veracity labels were associated with increased user interactions with the labeled tweets (likes by 53%, retweets by 85%, quote tweets by 120%, and replies by 76%, all statistically significant except for likes). Contextual labels were also associated with increased user interactions other than quote tweets, although the effect sizes were smaller as compared to veracity labels (likes 44%; retweets 65%; quote tweets: -3% replies: 18%) and the estimators were not statistically significant. Importantly, the design of labels (rebuttal strength and overlap) mitigated or even reversed these engagement effects. Where labels overlapped with the underlying tweet either topically or linguistically, there was a decrease of users interactions, with differences between the two types of textual overlap being negligible. On average, textual overlap corresponded to a decrease of 47% for likes, 40% for retweets, 45% for quote tweets and 39% for replies (statistically significant for likes, retweets, and replies). Labels with moderate or high rebuttal strength had no statistically significant association with users liking (moderate: +23% & strong: +39&), retweeting (moderate: + 7% & strong: +21%), quote tweeting (moderate: -25% & strong: +2%), and replying (moderate: -25%; strong: +22%). Much more robust than any label effect was the falsity of the underlying tweet. Trump's false statements always increased user interactions in all four ways, regardless of the topic of misinformation. Fraud (Ballot) related statements increased user liking by 68% (69%), retweeting by 68% (97%), replying by 66% (140%) and quote tweeting by 73% (130%). Arguments that Trump won the election attracted on average 52% more likes, 22% more retweets, 56% more quote tweets, and 43% more replies.

Overall, these results show that both in terms of misinformation spread (retweets, quote tweets) and user engagement (likes, replies), labels were not associated with increased user interactions. In fact, labels were associated with reduced interactions when tailored appropriately to the tweet content (in terms of textual overlap & rebuttal strength). The models also show that veracity labels were associated with more user interactions than were contextual labels. Furthermore, the association between false content and increased user engagement was larger than anything to do with tweet labels, provided these labels were germane to the underlying tweet. These findings answer a part of the RQ, namely how did different warning labels placed on Trump's false tweets impact (a) misinformation spread and (b) magnitude of user engagement.

Table 2: Median user (in terms of tweeting activity) and distribution of partisans replying to Trump's tweets containing misinformation across label types and properties.

	Median user (Nr. of tweets)	Liberal	Convervative
Without label	13	72%	28%
Veracity label	9	80%	20%
Contextual label	8	50%	50%
Labeled, without overlap	12	55%	45%
Topical overlap	10	58%	42%
Linguistic overlap	11	57%	43%
Labeled, without rebuttal	10	53%	47%
Moderate rebuttal	9	57%	43%
Strong rebuttal	11	57%	43%

4.3 Labels & type of engaging users

Besides understanding whether the placement of warning labels resulted in more or less misinformation sharing and user engagement, it is important to understand *who* engaged with the labeled tweets and *how*. Table 2 presents the distribution of users that responded to labeled and unlabeled tweets containing misinformation, broken down by level of tweeting activity and political orientation. Understanding who replied to Trump's tweets based on their tweeting activity level can shed light on how labels were associated with a mobilization or demobilization of the median (passive) Twitter user.

Comparing labeled and unlabeled tweets containing misinformation, labeled tweets engaged more active users in generating replies. A replying user had a median number of 13 tweets, compared to 9 for those responding to tweets with veracity labels and 8 for those responding to tweets with contextual labels (AD pair-wise tests statistically significant). In terms of textual overlap between label and tweet, tweets that had linguistic and topical overlap attracted more passive users to engage with them. A user replying to a labeled tweet without textual overlap had a median number of 12 tweets, while it was 10 and 11 for tweets with topical and linguistic overlap respectively (AD pair-wise tests statistically significant). In contrast, labels with strong rebuttal strength attracted more active users (11) compared to labeled tweets without any rebuttal strength, or a with a moderate one (10 and 9 respectively, AD pair-wise tests statistically significant). These results illustrate that different types of labels were associated with different user populations engaging with the tweets, even when controlling for tweet content. These preliminary findings show that in most cases, labels were associated with greater engagement (as measured by increased replying) by passive users, who are the majority of users on the platform.

In addition to user activity level, we looked at user political affiliation as a variable in engagement behavior. Table 2 illustrates how users across the partisan spectrum interacted with labeled and unlabeled tweets. Results show that unlabeled false tweets generally attracted the engagement, in terms of replies, of more

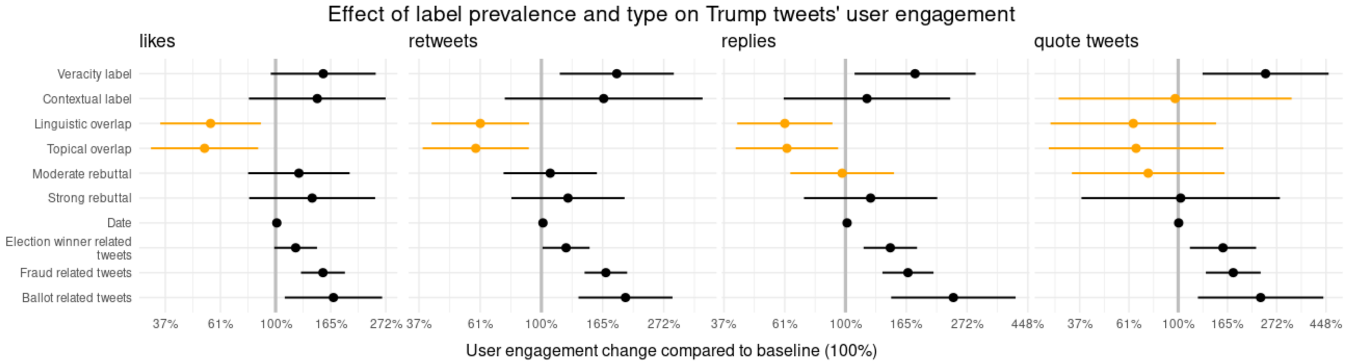


Figure 2: Forest plots of the robust linear regression models. For each interaction type (likes, retweets, replies, quote tweets), each plot shows the corresponding estimator effects. Confidence intervals at $a = 0.05$.

liberal users, with a difference of 44% compared to conservatives. This difference increased for tweets having a veracity label (60% more liberals), while it diminished for tweets containing contextual labels (50-50). The distribution of partisan users was different when it came to the existence of textual overlap between the label and underlying tweet content. Liberal users seemed to engage more with tweets that exhibited textual overlap (linguistic or topical) in the label than with those that did not (10% more liberals than conservatives without textual overlap, 16% with topical overlap and 14% with linguistic overlap). Focusing on labels' rebuttal strength, there were 6% more liberal users replying to tweets with labels of no rebuttal strength, and 14% more to tweets with moderate and strong rebuttals.

These results show that the type of user engagement, in terms of user activity and partisanship, varied depending on the type of label, when controlling for misinformation. Although the above results do not necessarily imply causation, they provide evidence that the choice of warning label properties could have an impact on *who* will engage, in terms of replies, with labeled content.

4.4 Labels & content of replies

The content analysis of the replies to Trump's unmoderated and soft- moderated tweets reveals additional associations between warning labels and user engagement type. The topic modeling algorithm yielded 64 distinct topics in the discussions under Trump's tweets (Appendix, Figure 5). Replies to unlabeled tweets included discussions of COVID-19 and taxation, for example. Replies to both labeled and unlabeled tweets included discussions of US foreign affairs, critiques of Trump(e.g., for losing the election and spreading lies), support for Trump, and US states and science. Replies that appeared over proportionately under labeled tweets were mostly related to election fraud. Overall the prevalence of these tweets was 8.5 % on labeled tweets compared to 5.5% on unlabeled ones.

We performed stance detection classification on all tweets belonging to these topics, and deployed a multinomial model to uncover explicit associations between warning labels and the content in user replies. The results in figure 3 (also appendix - table 5) illustrate that various properties of the labels influenced the generation of arguments supporting or rejecting election fraud narratives. Both veracity and contextual labels were associated with reduced chances that a user would discuss fraud, as opposed to talking about something else, regardless of the user's priors with respect to the misinformation. Contextual labels had a larger effect. By contrast, moderate and strong rebuttals were associated with greater chances that a user would discuss fraud (1-1.5 times increase in odds). Interestingly, textual overlap had an asymmetric effect on how users externalized their stances. It increased the likelihood of a reply rejecting Trump's assertions of fraud (circa 1.1 times), while decreasing the likelihood of a reply supporting the fraud theory (0.9 times). This suggests that textual overlap mobilized users whose opinion aligned with the content of the warnings. Focusing on partisanship and user activity level, liberal users were more likely, as expected, to reject misinformation in their replies, while conservative users were more likely to support it (for both p<0.01). More active users were also more likely to externalize their stance, albeit the difference was negligible (less than 1% change in odds, p<0.01).

The beta regression results (Figure 3- right, also appendix - table 5) also uncover associations between toxicity in replies, warning labels, and users. Both veracity and contextual labels increased the probability that users would generate toxic language, even when controlling for misinformation-related content (1.08 and 1.1 times increase in the odds respectively, p < 0.01). Textual overlap had no statistically significant effect on the amount of toxicity users generated. By contrast, the stronger the label's rebuttal, the less toxic were the replies (0.9 and 0.85 times for moderate and strong rebuttal). In terms of partisanship and user type, liberal users were more likely to create toxic replies in the discussions (1.03 times increase in the odds, p<0.01), as were the more passive users, although the difference was again negligible (less than 1% odds change, p<0.01).

The above findings answer the last part of the RQ, namely how warning labels interact with user engagement type. They show that warning label type corresponded to distinct distribution of reply topics related to Trump's tweets, even when controlling for tweet content. Some labels are associated with more users externalizing their stance on misinformation (pro or anti), while at the same time with users deploying less toxic language.

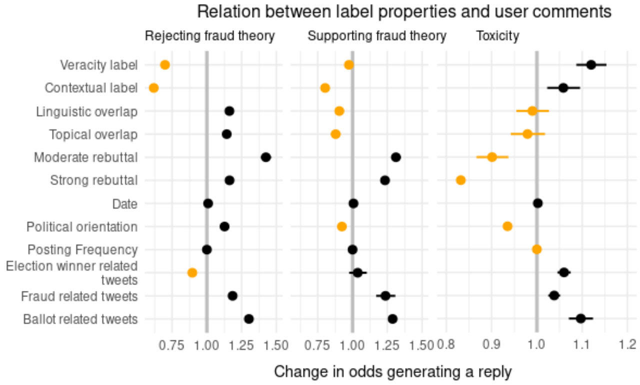


Figure 3: Left & Center: Forest plots of the factors that contribute in users creating arguments that either reject or support fraud theories, compared to other types of replies. Right: Forest plot of the factors that influence the propensity of a user generating a toxic reply. Each plot shows the corresponding estimator effects. Confidence intervals at $a = 0.05$.

5 DISCUSSION

Building on previous research findings [42, 57], the study investigated the initial positive relation between warning labels on Trump's tweets, corresponding misinformation spread, and user engagement. Controlling for tweet content, we found that there was actually no statistically significant difference in engagement between labeled and unlabeled tweets. Indeed, the most important driver for increased interactions was the presence of false tweet content (whether labeled or not). These results add to previous research studies showing that warning labels do not necessarily produce the intended results in terms of behavioral outcomes [23, 44]. Nevertheless, the fact that we did not find any backfire effects, as other studies have found [31, 33], supports the view that warning labels can inform the public about the truthfulness of social media content, as an alternative to hard moderation practices, without producing negative epistemic consequences.

Further results of our study show that different kinds of users may react differently to different kinds of labels. Properties of labels such as rebuttal strength and textual overlap were associated with less user interactions with false content. Although we looked at a concrete case study with a limited amount of data, these findings can be used to investigate more systematically how to design warning labels in the future depending on particular goals. Textual overlap and strong rebuttals are properties that attempt to yoke the tweet content to factual evidence. Many studies (e.g. [10, 36]) have shown that explicitly shifting the public's attention to accuracy, can have a mitigating effect on misinformation belief. Our study finds a similar association between an emphasis on labels' rebuttal strength and textual overlap with the underlying tweet and differences in user interactions, and hence shows a potential pathway for future research and policy interventions.

In addition, our study uncovered that warning labels can affect *who* generates content on social media and *how*. We analyzed the activity level and partisanship of the median user who replied to labeled tweets. Our results showed that labels with stronger rebuttals and textual overlap generally activated more liberal users to engage, while labels in general mobilized less active users, who are more typical of Twitter's user base. We found that stronger rebuttals in labels reduced toxicity in discussions and were more likely to make users externalize their opinion about misinformation. These results are valuable insofar as content soft-moderation seeks to influence how political and other discussions take place online. The fact that specific label properties are associated to with less toxicity should inform consideration of the value and structure of soft-moderation policies.

Our study shows that warning labels on misinformation may affect user behavior in complex ways beyond basic measures of user engagement. Depending on the characteristics of the warning label, there may be more user interactions, but of a less toxic nature; there may be more user engagement, but asymmetrically among partisans; there may be less engagement, but more among active than passive users. Further research is needed to systematize specific findings, and translate them into policy interventions.

6 CONCLUSION

In this study, we investigated the warning labels placed on Donald Trump's tweets about the 2020 US Presidential election. We categorized labels by type, rebuttal strength and textual overlap and deployed statistical tools to understand their relation to misinformation spread and user engagement. We found that, overall, warning labels did not change the magnitude of users interacting with tweets. Nevertheless, we found that the existence of textual overlap did reduce users' propensity to like, retweet, quote tweet, and reply to misinformation. Labels' rebuttal reduced users' propensity to create toxic content and made them externalize their stance more. We found that less active Twitter users tended to engage more with labeled tweets, depending on how the label was written, and that liberals mobilized more than conservatives in replies when content was labeled. These findings have direct implications for soft-moderation design.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Information Technology Policy at Princeton University. The authors also thank the Knight Foundation for research support. They thank Ben Kaiser and Eli Lucherini for feedback on the manuscript, Sanne Kruikemeier and Tom Dopper for fruitful discussions on the research process, Ashley Gorham for the support in the initial steps of the study, and Minos Papakyriakopoulos and Lev Cohen for their contributions in tweets' labeling. They also thank the anonymous reviewers for their constructive comments.

REFERENCES

[1] 2020. Additional steps we're taking ahead of the 2020 US Election. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes
[2] 2020. COVID-19 misleading information policy. https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy
[3] 2020. An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19. https://about.fb.com/news/2020/04/covid-19-misinfo-update/
[4] 2021. Shadow bans, fact-checks, info hubs: The big guide to how platforms are handling misinformation in 2021. https://www.niemanlab.org/2021/06/shadow-bans-fact-checks-info-hubs-the-big-guide-to-how-platforms-are-handling-misinformation-in-2021/?utm_source=DailyLabemaillist&utm_campaign=

75be0d870a-dailylabemail3&utm_medium=email&utm_term=0_d68264fd5e-75be0d870a-365017293

[5] 2021. Using machine learning to reduce toxicity online. https://perspectiveapi.com/

[6] Donald WK Andrews. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society* (1991), 817–858.

[7] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv preprint arXiv:2005.04682* (2020).

[8] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23, 1 (2015), 76–91.

[9] Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics.* Oxford University Press.

[10] Nadia M Brashier, Emmaline Drew Eliseev, and Elizabeth J Marsh. 2020. An initial accuracy focus prevents illusory truth. *Cognition* 194 (2020), 104054.

[11] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science* 28, 11 (2017), 1531–1546.

[12] Chris Cillizza. 2021. Analysis: 1 in 3 Americans believe the 'Big Lie'. https://www.cnn.com/2021/06/21/politics/biden-voter-fraud-big-lie-monmouth-poll/index.html

[13] Geoffrey L Cohen. 2003. Party over policy: The dominating impact of group influence on political beliefs. *Journal of personality and social psychology* 85, 5 (2003), 808.

[14] Ullrich KH Ecker, Stephan Lewandowsky, and Joe Apai. 2011. Terrorists brought down the plane!—No, actually it was a technical fault: Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology* 64, 2 (2011), 283–310.

[15] Ziv Epstein, Adam J Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G Rand. 2021. Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review* (2021).

[16] Don Fallis. 2015. What is disinformation? *Library trends* 63, 3 (2015), 401–426.

[17] Luciano Floridi. 2008. Semantic Conceptions of Information. *Stanford Encyclopedia of Philosophy* (2008).

[18] Aline Shakti Franzke, Anja Bechmann, Michael Zimmer, and Charles Ess. [n. d.]. the Association of Internet Researchers (2020). *Internet research: Ethical guidelines* 3 ([n. d.]).

[19] Kim Fridkin, Patrick J Kenney, and Amanda Wintersieck. 2015. Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising. *Political Communication* 32, 1 (2015), 127–151.

[20] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[21] Thomas T Hills. 2019. The dark side of information proliferation. *Perspectives on Psychological Science* 14, 3 (2019), 323–330.

[22] Hollyn M Johnson and Colleen M Seifert. 1994. Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of experimental psychology: Learning, memory, and cognition* 20, 6 (1994), 1420.

[23] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer. 2021. Adapting Security Warnings to Counter Online Disinformation. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 1163–1180. https://www.usenix.org/conference/usenixsecurity21/presentation/kaiser

[24] TaeYoung Kang and Jaeung Sim. 2021. Partisan Responses to Fact-Checking in Online News Platforms: Evidence from a Political Rumor about the North Korean Leader. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 266–277.

[25] Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4725–4735.

[26] Stephan Lewandowsky, John Cook, Nicolas Fay, and Gilles E Gignac. 2019. Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Memory & cognition* 47, 8 (2019), 1445–1456.

[27] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[28] Tony Liu, Lyle Ungar, and Konrad Kording. 2021. Quantifying causality in data science with quasi-experiments. *Nature Computational Science* 1, 1 (2021), 24–32.

[29] Ian Lundberg, Rebecca Johnson, and Brandon M Stewart. 2021. What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review* 86, 3 (2021), 532–565.

[30] Patricia L Moravec, Antino Kim, and Alan R Dennis. 2020. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research* 31, 3 (2020), 987–1006.

[31] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[32] Brendan Nyhan. 2021. Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences* 118, 15 (2021).

[33] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[34] Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. 2020. Political communication on social media: A tale of hyperactive users and bias in recommender systems. *Online Social Networks and Media* 15 (2020), 100058.

[35] Jessica Paynter, Sarah Luskin-Saxby, Deb Keen, Kathryn Fordyce, Grace Frost, Christine Imms, Scott Miller, David Trembath, Madonna Tucker, and Ullrich Ecker. 2019. Evaluation of a template for countering misinformation—Real-world Autism treatment myth debunking. *PloS one* 14, 1 (2019), e0210746.

[36] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[37] Gordon Pennycook and David G Rand. 2017. Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy. *Social Science Research Network. https://papers. ssrn. com/sol3/papers. cfm* (2017).

[38] Charles S Reichardt. 2009. Quasi-experimental design. *The SAGE handbook of quantitative methods in psychology* 46, 71 (2009), 490–500.

[39] Michael D Rich et al. 2018. *Truth decay: An initial exploration of the diminishing role of facts and analysis in American public life.* Rand Corporation.

[40] Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. Stm: An R package for structural topic models. *Journal of Statistical Software* 91, 1 (2019), 1–40.

[41] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[42] Zeve Sanderson, Megan A Brown, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2021. Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review* (2021).

[43] David Schkade, Cass R Sunstein, and Reid Hastie. 2010. When deliberation produces extremism. *Critical Review* 22, 2-3 (2010), 227–252.

[44] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. 2021. Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes. *arXiv preprint arXiv:2104.00779* (2021).

[45] Todd Spangler. 2020. TikTok Will Tag Election-Related Videos With Link to 2020 U.S. Voters Guide. https://variety.com/2020/digital/news/tiktok-elections-voting-guide-misinformation-trump-1234786313/

[46] Christina Steindl, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg. 2015. Understanding psychological reactance. *Zeitschrift für Psychologie* (2015).

[47] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition* (2020).

[48] Emily Thorson. 2016. Belief echoes: The persistent effects of corrected misinformation. *Political Communication* 33, 3 (2016), 460–480.

[49] Kerrie L Unsworth and Kelly S Fielding. 2014. It's political: How the salience of one's political identity changes climate change beliefs and policy support. *Global Environmental Change* 27 (2014), 131–137.

[50] Trevor Van Mierlo. 2014. The 1% rule in four digital health social networks: an observational study. *Journal of medical Internet research* 16, 2 (2014), e2966.

[51] Nathan Walter, Jonathan Cohen, R Lance Holbert, and Yasmin Morag. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication* 37, 3 (2020), 350–375.

[52] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe* 27 (2017).

[53] Chloe Wittenberg, Adam J Berinsky, Nathaniel Persily, and Joshua A Tucker. 2020. Misinformation and its correction. *Social Media and Democracy: The State of the Field, Prospects for Reform* 163 (2020).

[54] Stefan Wojcik and Adam Hughes. 2021. How Twitter Users Compare to the General Public. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

[55] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–14.

[56] Chun Yu and Weixin Yao. 2017. Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation* 46, 8 (2017), 6261–6282.
[57] Savvas Zannettou. 2021. " I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 865–876.
[58] Yanmengqian Zhou and Lijiang Shen. 2021. Confirmation Bias and the Persistence of Misinformation on Climate Change. *Communication Research* (2021), 00936502211028049.

A ETHICAL & REPRODUCIBILITY CONCERNS

No private information has been collected or processed during this study. All data are public data, as given to the researchers following the Twitter non-commercial developer agreement. Nevertheless, to respect individuals' privacy and following the Association of Internet Researchers guidelines [18], the data has been completely deleted upon publication of the study, while all necessary anonymized materials to replicate findings can be found here: https://github.com/civicmachines/Twitter_labels_trump/.

B FIGURES

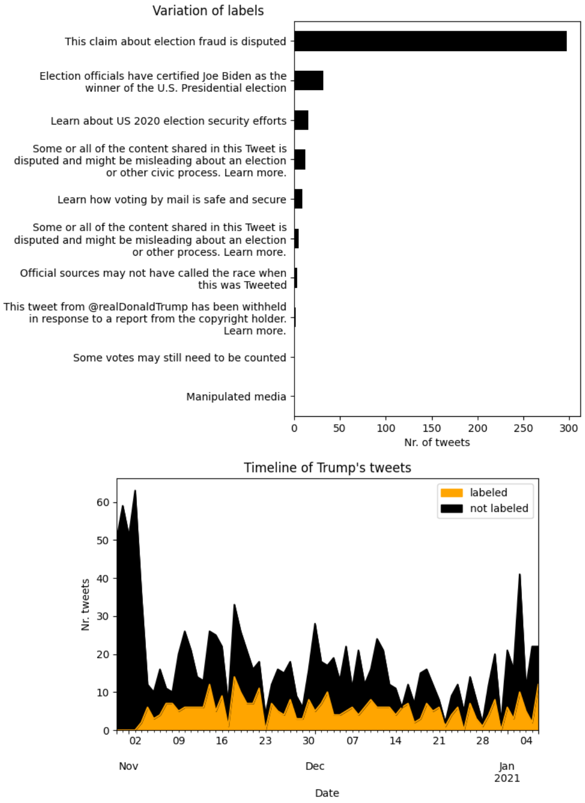


Figure 4: Up: Overview of the found warning labels and their prevalence on Trump's tweets (including retweets & quote tweets). Down: Timeline of labeling Trump's tweets

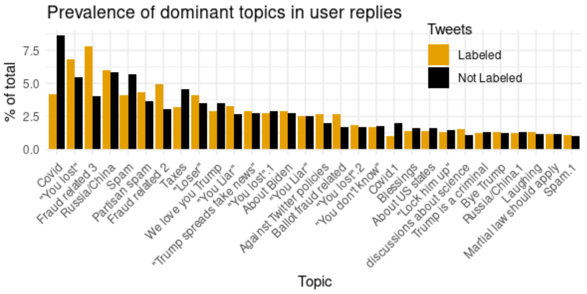


Figure 5: Prevalence of dominant topics in user replies. For labeled and unlabeled tweets, the figure present the percental share of replies belonging to each topic.

C TABLES

Table 3: Examples of the classification scheme used to characterize warning labels. V: Veracity label, C: Contextual label, RS: Rebuttal strength (1: moderate rebuttal, 2: strong rebuttal), LO: linguistic overlap, TO: topical overlap.

Tweet	Label	V	C	RS	LO	TO
RIGGED ELECTION. WE WILL WIN!	This claim about election fraud is disputed	1	0	1	0	1
I won the Election!	Election officials have certified Joe Biden as the winner of the U.S. Presidential election	1	0	2	1	0
We won Wisconsin big. They rigged the vote!	Learn about US 2020 election security efforts.	0	1	0	0	1
"Data group exposes wide spread Mail-In Ballot Fraud." @ChanelRion @OANN	Learn how voting by mail is safe and secure	0	1	1	1	0

Table 4: Robust linear regression results for the association of label properties and tweet content with user interactions (likes, retweets, replies, quote tweets)

	Likes	Retweets	Replies	Quote tweets
(Intercept)	16.93^{***}	14.50^{***}	9.41^{***}	9.79^{***}
	(0.13)	(0.11)	(0.13)	(0.15)
Veracity	0.43	0.62^{**}	0.57^{*}	0.89^{**}
	(0.24)	(0.24)	(0.25)	(0.33)
Contextual	0.38	0.51	0.18	−0.03
	(0.32)	(0.41)	(0.35)	(0.60)
Topical overlap	-0.59^{*}	-0.50^{*}	-0.50^{*}	−0.46
	(0.23)	(0.20)	(0.20)	(0.43)
Linguistic overlap	-0.64^{**}	-0.54^{*}	-0.48^{*}	−0.43
	(0.25)	(0.22)	(0.21)	(0.45)
Moderate rebuttal	0.21	0.07	−0.03	−0.30
	(0.23)	(0.19)	(0.22)	(0.39)
Strong rebuttal	0.33	0.22	0.21	0.03
	(0.29)	(0.24)	(0.28)	(0.51)
Date	0.01^{***}	0.01^{***}	0.01^{***}	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Winner	0.18	0.20^{*}	0.37^{**}	0.45^{**}
	(0.10)	(0.10)	(0.11)	(0.17)
Fraud	0.43^{***}	0.53^{***}	0.51^{***}	0.56^{***}
	(0.10)	(0.09)	(0.11)	(0.14)
Ballots	0.53^{*}	0.69^{***}	0.89^{***}	0.84^{**}
	(0.22)	(0.20)	(0.26)	(0.32)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5: Summary statistics for beta and multinomial regression results for the association of label properties with the toxicity in user replies and stance on fraud theories.)

	Supporting fraud theory	Rejecting fraud theory	Toxicity
(Intercept)	-3.07^{***}	-1.57^{***}	-0.44^{***}
	(0.03)	(0.02)	(0.00)
Veracity	-0.36^{***}	-0.03^{*}	0.11^{***}
	(0.01)	(0.01)	(0.02)
Contextual	-0.48^{***}	-0.22^{***}	0.06^{**}
	(0.00)	(0.00)	(0.02)
Topical overlap	0.15^{***}	-0.10^{***}	−0.01
	(0.01)	(0.02)	(0.02)
Linguistic overlap	0.13^{***}	-0.13^{***}	−0.02
	(0.00)	(0.02)	(0.02)
Moderate rebuttal	0.35^{***}	0.27^{***}	-0.10^{***}
	(0.01)	(0.01)	(0.02)
Strong rebuttal	0.15^{***}	0.21^{***}	-0.18^{***}
	(0.00)	(0.01)	(0.02)
Date	0.01^{***}	0.01^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)
Political orientation	0.12^{***}	-0.08^{***}	-0.07^{***}
	(0.01)	(0.01)	(0.00)
Nr. of user replies	0.00^{***}	0.00^{***}	-0.00^{***}
	(0.00)	(0.00)	(0.00)
Winner	-0.11^{***}	0.04	0.06^{***}
	(0.00)	(0.03)	(0.01)
Fraud	0.17^{***}	0.21^{***}	0.04^{***}
	(0.01)	(0.03)	(0.01)
Ballots	0.26^{***}	0.25^{***}	0.09^{***}
	(0.00)	(0.00)	(0.01)
Precision: (phi)			2.30^{***}
			(0.01)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$