"I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter

Savvas Zannettou

Max Planck Institute for Informatics
szannett@mpi-inf.mpg.de

Abstract

Over the past few years, there is a heated debate and serious public concerns regarding online content moderation, censorship, and the principle of free speech on the Web. To ease these concerns, social media platforms like Twitter and Facebook refined their content moderation systems to support soft moderation interventions. Soft moderation interventions refer to warning labels attached to potentially questionable or harmful content to inform other users about the content and its nature while the content remains accessible, hence alleviating concerns related to censorship and free speech.

In this work, we perform one of the first empirical studies on soft moderation interventions on Twitter. Using a mixed-methods approach, we study the users who share tweets with warning labels on Twitter and their political leaning, the engagement that these tweets receive, and how users interact with tweets that have warning labels. Among other things, we find that 72% of the tweets with warning labels are shared by Republicans, while only 11% are shared by Democrats. By analyzing content engagement, we find that tweets with warning labels had more engagement compared to tweets without warning labels. Also, we qualitatively analyze how users interact with content that has warning labels finding that the most popular interactions are related to further debunking false claims, mocking the author or content of the disputed tweet, and further reinforcing or resharing false claims. Finally, we describe concrete examples of inconsistencies, such as warning labels that are incorrectly added or warning labels that are not added on tweets despite sharing questionable and potentially harmful information.

Introduction

Social media platforms like Twitter and Facebook are under pressure from the public to address issues related to the spread of harmful content like hate speech (Guynn 2016) and online misinformation (Solon 2016), particularly during major events like elections. To ease the public's concerns and mitigate the effects of these important issues, platforms are continuously refining their guidelines and improving their content moderation systems (Hutchinson 2020).

Designing and implementing an ideal content moderation system is not straightforward as there are many challenges and aspects to be considered (Gillespie 2018). First, content
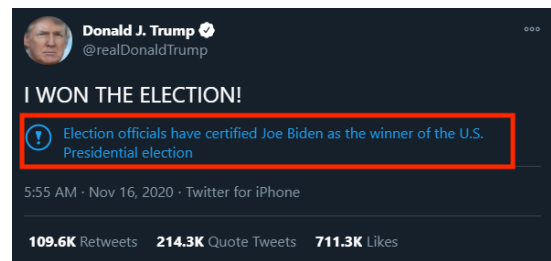
Figure 1: An example of a soft moderation intervention on Twitter.

moderation should be performed in a timely manner to ensure that harmful content is removed fast and only a small number of users are exposed to harmful content. This is a tough challenge given the scale of modern social media platforms like Twitter and Facebook. Second, content moderation should be consistent and fair across users. Finally, content moderation should ensure that it is in accordance with basic principles of our society like the freedom of speech.

To ease concerns related to freedom of speech and censorship, recently, Facebook and Twitter introduced a new feature in their content moderation systems; a type of soft moderation intervention that attaches warning labels and relevant information to content that is questionable or potentially harmful or misleading (Roth and Pickles 2020; Rosen 2020). An example of a soft moderation intervention is depicted in Fig. 1, where Twitter moderators attached a warning label to a tweet from President Trump related to the outcome of the 2020 US elections. These warning labels are designed to "correct" the tweet's content and provide necessary related information while ensuring that the freedom of speech principle is not violated.

Previous work investigated how users perceive these warning labels (Mena 2020; Geeng et al. 2020; Saltz, Leibowicz, and Wardle 2020; Seo, Xiong, and Lee 2019), assessing their effectiveness, how their design can affect their effectiveness (Bode and Vraga 2015; Kaiser et al. 2020; Moravec, Kim, and Dennis 2020), and possible unintended consequences from the use of warning labels (Pennycook et al. 2020; Nyhan and Reifler 2010). Despite this rich research work, the majority of it investigates these warning

labels in artificial environments either through interviews, surveys, or crowdsourcing studies. While these studies are useful and important, they do not consider platform-specific affordances such as user interactions with posts that have warning labels (e.g., retweets, likes, etc.) As a research community, we lack empirical evidence to understand how these warning labels are used on social media platforms like Twitter and how users interact and engage with them.

In this work, we aim to bridge this research gap by performing an empirical analysis of soft moderation interventions on Twitter. We focus on answering the following research questions:

- **RQ1:** What are the various warning labels used on Twitter during the 2020 US elections and what kind of users have their tweets flagged more frequently? Are there differences across political leanings?
- **RQ2:** Is the engagement of content that includes warning labels significantly different from the content without warning labels?
- **RQ3:** How do users on Twitter interact with content that includes warning labels?

To answer these research questions, we collect a dataset of tweets shared between March 2020 and December 2020, including soft moderation interventions (i.e., warning labels). To do this, we use Twitter's API and collect the timeline of popular verified users. We mainly focus on verified users as they usually have a large audience and their content can receive considerable engagement. Overall, we collect a set of 18K tweets that either had warning labels on them or quote a tweet that had a warning label. These tweets were shared from 8.1K users between March 2020 and December 2020. Due to our data collection method, 96% of the tweets in our dataset were shared between November 2020 and December 2020, hence our dataset is related to the 2020 US elections. Then, we follow a mixed-methods approach to analyze the engagement of tweets with warning labels and the users that share them (quantitative analysis), as well as how users interact with tweets and warning labels (qualitative analysis).

Findings. Our main findings are:
- We find that 72.8% of the tweets that include warning labels were shared by Republicans, while only 11.6% of the tweets were shared by Democrats. This likely indicates that during the 2020 US elections period, Republicans tended to disseminate more questionable or potentially harmful information (e.g., claims about election fraud) that is eventually flagged by Twitter (**RQ1**).
- By analyzing the engagement of tweets, we find that tweets that have warning labels receive more engagement compared to tweets without warning labels. Also, by looking into the users that have increased engagement in tweets with warning labels, we find that most of the users that have high engagement, in general, have increased engagement on tweets with warning labels as well (**RQ2**). Note that our analysis does not consider *when* the warning label was placed and the engagement before the addition of the warning label. Therefore, it is likely that the increased engagement on tweets with warning labels is because Twitter is prioritizing high engagement content in their soft moderation systems (Roth and Pickles 2020). In

other words, it is likely that the warning label was placed because the tweet already received high engagement.
- Our qualitative analysis indicates that a lot of users interact with content that has warning labels by further debunking false claims, mocking or sharing emotions about the author/content of the questionable tweet, or by reinforcing the false claims that are included in tweets with warning labels. Also, we shed light on some of the challenges and issues that exist when designing and developing large-scale soft moderation intervention systems. We find instances where the warning labels were incorrectly added (e.g., see Fig. 7) and cases where the moderation system is inconsistent (i.e., content should be flagged, but it is not). Some of these cases are likely due to the dissemination of similar information across different languages (e.g., see Fig. 8) and across various formats of information like text and videos (**RQ3**).

Contributions. The contributions of this work are three-fold. First, to the best of our knowledge, we perform one of the first characterizations of soft moderation interventions based on empirical data from Twitter. Also, we make our dataset publicly available (Zannettou 2021), hence assisting the research community in conducting further studies on soft moderation interventions based on empirical data. Second, our quantitative analysis sheds light on the engagement that content with soft moderation interventions receive on Twitter. This analysis encapsulates engagement from real users interacting with timely content on Twitter, hence it complements and strengths the findings from studies undertaken in controlled experiments (e.g., via surveys). Finally, our qualitative analysis highlights how users interact with content that includes warning labels, and it helps us understand some of the real-world challenges that exist when designing soft moderation intervention systems.

Background & Related Work

Moderation interventions on social media platforms can be applied on various levels. First, some interventions are applied at the post level (e.g., post-removal). Second, some interventions exist on the user level (Myers West 2018; Merrer, Morgan, and Trédan 2020) like user bans or shadow banning (i.e., limiting the visibility of their activity). Finally, community-wide moderation interventions exist where platforms moderate specific sub-communities within their platforms (e.g., banning Facebook groups or subreddits) (Chandrasekharan et al. 2020, 2017; Newell et al. 2016; Ribeiro et al. 2020; Saleem and Ruths 2018).

For each of the above-mentioned levels, there are two different types of interventions: *hard* and *soft* interventions. Hard moderation interventions refer to moderation actions that remove content or entities from social media platforms (posts, users, or communities). On the other hand, soft moderation interventions do not remove any content, and they aim to inform other users about potential issues with the content (e.g., by adding warning labels) or limiting the visibility of questionable content (shadow banning) or imposing restrictions on the ability of other users to engage with the content (i.e., users are unable to reply, re-share, etc.). Below,

we review relevant previous work that studies post-level soft moderation interventions as they are the most relevant to our work.

Post-level Soft Moderation Interventions

A rich body of previous work investigates soft moderation interventions mainly through interviews, surveys, and crowdsourcing studies. Specifically, (Mena 2020) performs an experiment using Amazon Mechanical Turk (AMT) workers to understand user perceptions on content that includes warning labels. By recruiting Facebook users and performing crowdsourcing studies, they find that the warning label had a significant effect on users' sharing intentions; that is, participants were less willing to share content with warning labels. (Geeng et al. 2020) focus on warning labels added on Twitter, Facebook, and Instagram related to COVID-19 misinformation. Through surveys, they find that users have a positive attitude towards warning labels. However, they highlight that users verify misinformation through other means and search on the Web for relevant information. (Saltz, Leibowicz, and Wardle 2020) focus on warning labels added on visual misinformation related to COVID-19. By conducting in-depth interviews, they find that participants had different opinions regarding warning labels, with many participants perceiving them as politically-biased and an act of censorship from the platforms.

(Kaiser et al. 2020) use methods from information security to evaluate the effectiveness and the design of warning labels. Through controlled experiments, they find that despite the existence of warning labels, users seek information via other means, thus confirming the findings from (Geeng et al. 2020). Also, by performing crowdsourcing studies and asking users about 8 warning label designs, they conclude that users' information-seeking behavior is significantly affected by the design of the warning label. (Seo, Xiong, and Lee 2019) investigate user perceptions when they are exposed to fact-checking and machine learning-generated warning labels. Through experiments on AMT, they find that users trust more fact-checking warning labels than machine learning-generated ones. (Moravec, Kim, and Dennis 2020) highlight that the design of warning labels (i.e., how warnings are presented to users) can change their effectiveness. Also, they emphasize that clearly explaining the warning labels to users can lead to increased effectiveness. (Bode and Vraga 2015) study the related stories functionality on Facebook as a means to detect or debunk misinformation. By conducting surveys, they find that when related stories debunk a misinformation story, it significantly reduces the participants' misperceptions (beliefs that are not supported by evidence or expert opinion (Nyhan and Reifler 2010)).

Other previous work demonstrates some unintentional consequences from the use of warning labels. Specifically, (Pennycook et al. 2020) conduct AMT studies and they show an *implied truth* effect, where posts that include misinformation and are not accompanied with a warning label are considered credible. Also, (Nyhan and Reifler 2010) conduct controlled experiments to assess the effectiveness of warning labels to political false information. They highlight a *backfire effect*, where participants strengthen their support to false political stories after seeing the warning label that includes a correction. Follow-up research on the backfire effect (Wood and Porter 2019) was unable to find instances of it, while another work (Swire-Thompson, DeGutis, and Lazer 2020) suggests that this effect is highly dependent on context. Finally, (Pennycook, Cannon, and Rand 2018) emphasize the existence of the *illusory truth effect* where users tend to believe false information after getting exposed to it multiple times or for an extended time period, even though the false information is accompanied by a warning label.

Remarks. Previous work investigated soft moderation interventions in artificial testing environments like interviews, surveys, and crowdsourcing studies. This previous work is essential as it helps us understand how people intend to interact and engage with content that includes warning labels or corrections. However, they do not capture platform-specific peculiarities, and they do not adequately capture how people interact and engage with warning labels in realistic scenarios (e.g., when reading a tweet from the US President during the US elections period). In this work, we address these limitations by performing one of the first empirical analysis of soft moderation interventions on Twitter.

Dataset

We start our data collection on Twitter and, in particular, on verified users, which are users who have an "especially large audience or are notable in government, news, entertainment, or another designated category" (the definition is obtained from the Twitter website). We mainly focus on verified users as they usually have a large audience and can have a substantial impact on online discussions, hence moderating content from these users is important.

We collect the dataset of Twitter verified users from Pushshift (Pushshift 2020). The dataset includes Twitter account metadata for 351,655 verified users. Then, for each user, we use Twitter's API to obtain recent tweets/retweets shared by these users (i.e., their timeline). We also collect soft moderation-specific metadata for each tweet: these include whether a tweet is accompanied by a warning label and relevant metadata (e.g., label text). Note that we only collect warning labels placed below the tweets and not any other types of interventions like tweets placed below an interstitial like "This tweet may include sensitive content" (Twitter 2020). Due to the rate-limiting of the Twitter API, we tried to collect activity only from the top 170,506 users based on the number of their followers (corresponding to 48.4% of all the Twitter verified accounts in the Pushshift dataset). We managed to collect data for 168,126 users, as the rest were either deleted, suspended, or accounts set to private. Our dataset collection process was conducted between December 7, 2020 and December 31, 2020. Overall, we collect 79,361,081 tweets shared during 2020 from 168,126 users.

Next, we select all tweets with soft moderation interventions (i.e., warning labels) from our dataset; we find 29,232 tweets from 9,334 verified users. This dataset also includes retweets of tweets with warning labels and tweets that quote a tweet with a warning label. Due to this, we rehydrate, using the Twitter API, all quoted and retweeted tweets that had a

	#Tweets	#Users
Tweets with warning labels (Fig. 1)	2,244	853
Quoted-Warning on quoted (Fig. 9)	16,571	7,651
Quoted-Warning on comment (Fig. 8)	219	98
Quoted-Warning on both (Fig. 7)	50	30
Total	18,765	8,142

Table 1: Overview of our dataset.

Warning label	%Tweets
1. This claim about election fraud is disputed	58.1%
2. Learn about US 2020 election security efforts	12.1%
3. Manipulated media	8.7%
4. Learn how voting by mail is safe and secure	5.8%
5. Official sources may not have called the race when this was Tweeted	4.5%
6. Multiple sources called this election differently	4.2%
7. Election officials have certified Joe Biden as the winner of the U.S. Presidential election	2.8%
8. Some votes may still need to be counted	1.1%
9. Get the facts about COVID-19	1.1%
10. Esta reivindicação de fraude é contestada	0.5%
11. Saiba por que urnas eletrônicas são seguras	0.5%
12. Sources called this election differently	0.1%
13. Get the facts about mail-in ballots	0.1%

Table 2: Warning labels in our dataset (the numbers are for illustration purposes and are not part of the warning label).

warning label; we get an additional 3,106 tweets from 1,888 users. Note that this procedure resulted in the acquisition of tweets from unverified users, as verified users in our dataset retweeted or quoted tweets from unverified users. Given that this content appears on the followers of verified users, we keep in our dataset tweets from unverified users.

After excluding all retweets, our final dataset consists of 18,765 tweets that include warning labels (either on the tweet itself or on quoted tweets) from 8,142 users (see Table 1). We split our dataset into two parts; 1) tweets that have warning labels attached to them (first row in Table 1); and 2) tweets that quote other tweets and any (or both) of the tweets have warning labels (see second-fourth row in Table 1). Note that there is overlap on the tweets with warning labels and quoted tweets for the cases where the warning label was on the comment above. For the remainder of this paper, we call the first part of our dataset *tweets with warning labels* and the second part of our dataset *quoted tweets*. Due to our data collection approach, a large percentage of the tweets in our dataset (96%) are shared between November 1, 2020 and December 30, 2020, hence our dataset is related to the 2020 US elections and has a political nature.

Ethical Considerations and Data Availability. We emphasize that we collect and work entirely with publicly available data as we do not collect any data from users who have a private account. Overall, we follow standard ethical research standards (Rivers and Lewis 2014) like refraining from tracking users across sites. To help advance empirical research related to soft moderation interventions on Twitter, we make publicly available (Zannettou 2021) the tweet IDs and their corresponding warning labels.

RQ1: Warning Labels and Users

In this section, we analyze the different types of warning labels and how they appear over time on Twitter. Also, we perform a user-based analysis aiming to uncover differences across users that have opposing political leaning.

Warning Labels

We start by looking into the different types of warning labels that exist in our dataset. To do this, we focus on tweets that include warning labels (see the first row in Table 1), specifically, 2,244 tweets posted by 853 users.

Table 2 shows all warning labels in our dataset and their respective frequency and percentage over all the tweets. Overall, we find 13 different warning labels with the majority of them being related to the 2020 US elections. For

instance, the most popular warning label in our dataset is "This claim about election fraud is disputed" with 58% of all tweets. Other 2020 US election warning labels are related to the security of the elections like "Learn about US 2020 election security efforts" (12%) and "Learn how voting by mail is safe and secure" (5.8%), as well as related to the outcome of the elections like "Multiple sources called this election differently" (4.2%) and "Election officials have certified Joe Biden as the winner of the U.S. Presidential election" (2.8%). Interestingly, we also find warning labels referring to the 2020 US elections written in other languages (i.e., Portuguese). We find 0.49% tweets including "Esta reivindicação de fraude é contestada" (translates to "This fraud claim is disputed") and "Saiba por que urnas eletrônicas são seguras" (translates to "Find out why electronic voting machines are safe"). We manually examine the tweets that had warning labels in Portuguese, finding that all tweets were posted in the Portuguese language by 15 different users (the language of the warning label is independent of the user's language preference on Twitter). Apart from politics-related warning labels, we find a general-purpose warning label that informs users about manipulated media (images or videos) with 8.7% of all tweets in our dataset. Finally, we find a COVID-19 specific warning label: "Get the facts about COVID-19" (1.15%) that aims to inform users about health-related issues and, in particular, the COVID-19 pandemic. The small percentage of tweets with COVID-19 related warning labels is likely because our dataset is mainly centered around the time period of the 2020 US elections.

Next, we analyze how these warning labels appear over time on Twitter. Fig. 2 shows how the top 10 most popular warning labels in our dataset appear over time on Twitter. We focus on the period between November 2020 and December 2020 for readability purposes and because most of the tweets in our dataset are shared during this period. We find two different temporal patterns. First, we find warning labels that are short-lived as most of their appearances on tweets happen within a short period of time. Concretely, both "Learn about US 2020 election security efforts" (#2) and
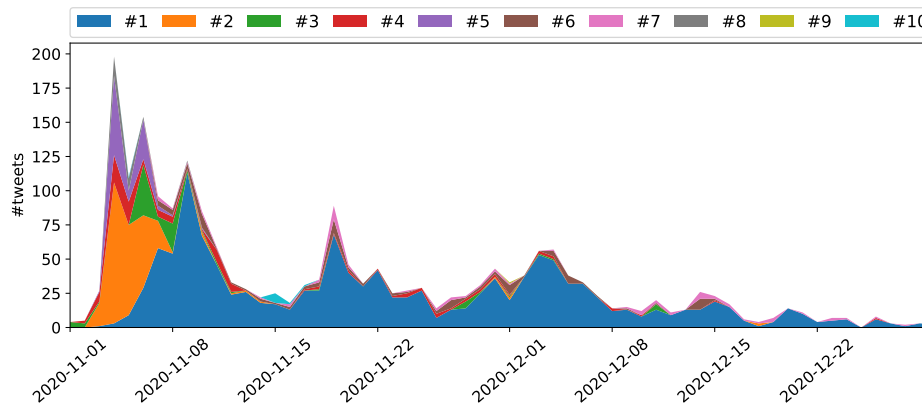
Figure 2: Number of tweets that include a warning label for each day between November 1, 2020 and December 30, 2020 (for the top 10 warning labels). See Table 2 for mapping the warning label text to the numbers in the legend.
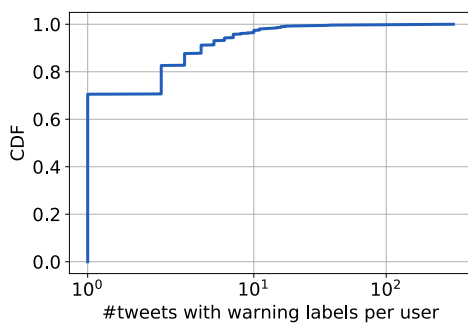

Figure 3: Cumulative Distribution Function (CDF) of the number of tweets with warning labels per user.

"Official sources may not have called the race when this was Tweeted" (#5) are exclusively used during the first week of November 2020. On the other hand, we find warning labels that are long-lived. E.g., the label "This claim about election fraud is disputed" (#1) is used for the entirety of the period between November and December 2020. Another example is "Manipulated media" (#3) used during the entire period (from March 2020 to December 2020). Overall, these results indicate that warning labels are time and context-dependent, with some of them being short-lived (few days) and some of them being long-lived (several months).

User Analysis

Here, we look into the users who share tweets with warning labels. Recall that our data collection involves 168K users, and only 853 of them share tweets that have warning labels, hence indicating that only a small percentage (0.5%) of Twitter verified users have warning labels attached to their content. As per Fig. 3, out of the 853 users, 70% had only one tweet with a warning label, while only 3.6% of these users had at least 10 tweets with warning labels. Overall, only a small percentage of users have warning labels on multiple of their tweets.

Users' Political Leaning. As we described above, our dataset has a strong political nature, and the majority of the warning labels refer to claims about the 2020 US elections (e.g., claims about election fraud, see Table 2). Motivated by this, we augment our dataset with information about the political leaning of each user that shared tweets with warning labels. To infer users' political leaning, we use the methodology presented by (Kulshrestha et al. 2017) and, in particular, the Political Bias Inference API that is publicly available by (Messias 2017). The API generates a vector with the topical interests of each user and their frequency. To do this, the API collects all the user's friends (i.e., people that the user follows), generates all the topics inferred for each friend using the methodology in (Ghosh et al. 2012; Sharma et al. 2012), and calculates a vector with all the topics and their frequencies. Finally, by comparing the topical vectors to a ground truth dataset of Republican and Democrat Twitter users, the API infers whether a Twitter user has a Republican, Democrat, or Neutral political leaning.

We use the Political Bias Inference API, between January 3 and January 10, 2021, to infer the political leaning of the 8,142 Twitter users in our dataset. We quantify the performance of the Political Bias Inference API in our dataset by manually classifying 200 users, finding 80.5% accuracy (see Appendix). Table 3 reports the number of tweets and users per inferred political leaning for the entire dataset and broken down into tweets with warning labels and quoted tweets. We observe that for the entire dataset, 51% of the users are Democrats, 13.4% are Republicans, almost 32% are inferred as neutral, while for the rest 1.4% we were unable to infer their political leaning. This is because some users were either suspended or made their accounts private by the time we were collecting their friend list, hence the Political Bias Inference API was unable to make an inference.

Interestingly, when looking at the tweets with warning labels in Table 3, we find that most of the tweets with warning labels are shared by Republicans (72% of all tweets vs. 11% for Democrats). This likely indicates that due to the context and developments related to the 2020 US elections, Republicans tend to share more questionable content that is more likely to receive warning labels by Twitter. For the

	All		Tweets with warning labels		Quoted tweets	
	Tweets	**Users**	**Tweets**	**Users**	**Tweets**	**Users**
Republicans	4,232 (22.5%)	1,095 (13.4%)	**1,634 (72.8%)**	**404 (47.3%)**	2,775 (16.5%)	917 (11.9%)
Neutral	4,610 (24.5%)	2,604 (31.9%)	294 (13.1%)	184 (21.5%)	4,347 (25.9%)	2,483 (32.3%)
Democrats	9,747 (51.9%)	4,324 (53.1%)	262 (11.6%)	218 (25.5%)	**9,491 (56.6%)**	**4,198 (54.7%)**
N/A	176 (0.9%)	119 (1.4%)	54 (2.4%)	47 (5.5%)	127 (0.7%)	75 (0.9%)

Table 3: Inferred political leaning of users who shared tweets with warning labels or quoted a tweet that had a warning label.

User	Political Leaning	Account Status	Tweets
realDonaldTrump	Republican	Suspended	321 (14.3%)
TeamTrump	Republican	Suspended	105 (4.6%)
gatewaypundit	Republican	Active	71 (3.1%)
va_shiva	Neutral	Active	38 (1.6%)
JudicialWatch	Republican	Active	36 (1.6%)
MichaelCoudrey	Republican	Suspended	27 (1.2%)
TomFitton	Republican	Active	20 (0.8%)
RudyGiuliani	Republican	Active	17 (0.7%)
JamesOKeefeIII	Republican	Active	17 (0.7%)
EmeraldRobinson	Republican	Active	16 (0.7%)
RealJamesWoods	Republican	Active	16 (0.7%)
LLinWood	Republican	Suspended	16 (0.7%)
realLizUSA	Republican	Active	15 (0.6%)
LouDobbs	Republican	Active	15 (0.6%)
KMCRadio	Republican	Suspended	14 (0.6%)
michellemalkin	Republican	Active	13 (0.5%)
CodeMonkeyZ	Republican	Suspended	12 (0.5%)
charliekirk11	Neutral	Active	11 (0.4%)
TrumpWarRoom	Republican	Active	11 (0.4%)
chuckwoolery	Republican	Active	11 (0.4%)

Table 4: Top 20 users who had the most warning labels on their tweets (users with bold username are unverified users). We also report the account status of each user as of January 9, 2021. We do not anonymize any username because all users are public figures (including the unverified users).

User	Political Leaning	Account Status	Tweets
realDonaldTrump	Republican	Suspended	78 (0.4%)
AndrewFeinberg	Democrat	Active	52 (0.3%)
svdate	Neutral	Active	50 (0.3%)
NumbersMuncher	Republican	Active	49 (0.3%)
atrupar	Democrat	Active	44 (0.2%)
GlennKesslerWP	Democrat	Active	42 (0.2%)
T_S_P_O_O_K_Y	Republican	Active	42 (0.2%)
BrianKarem	Democrat	Active	39 (0.2%)
Patterico	Republican	Active	38 (0.2%)
TalbertSwan	Neutral	Active	37 (0.2%)
BarnettforAZ	Republican	Active	33 (0.2%)
TomFitton	Republican	Active	32 (0.2%)
Justin_Stangel	Democrat	Active	31 (0.2%)
JLMarchese111	N/A	Suspended	31 (0.2%)
HalSparks	Democrat	Active	30 (0.2%)
RhondaFurin	Republican	Active	29 (0.2%)
captainjanks	Democrat	Active	28 (0.2%)
rogerkimball	Republican	Active	27 (0.2%)
amhfarraj	Democrat	Active	26 (0.2%)
michellemalkin	Republican	Active	25 (0.1%)

Table 5: Top 20 users who quoted tweets that had warning labels. We also report the account status of each user as of January 9, 2021.

quoted tweets, we observe that Democrats tend to comment on tweets with warning labels more often than Republicans (56% vs. 16.5% for Republicans).

Top Users. But who are the users who are the most "prolific" with regards to tweets that include warning labels or in the quoted tweets? Table 4 and Table 5 show the top 20 users in our dataset based on the number of tweets with warning labels and the quoted tweets, respectively. For each user, we report the inferred political leaning and whether the account was active or suspended on January 9, 2021. We make several observations. First, in both cases, the most prolific user is President Trump, with 14.3% of all tweets with warning labels and 0.4% of all quoted tweets. The account of President Trump was permanently suspended by Twitter on January 8, 2021, due to the risk of further incitement of violence (Twitter Inc. 2021), after his supporters attacked the US capitol (Evelyn 2021). Second, we observe that most of the top 20 users who shared tweets with warning labels are inferred as Republicans (see Table 4). This is not the case for

the quoted dataset (see Table 5), as 8 out of the top 20 users with quoted tweets are inferred as Democrats. Third, even though our study does not focus on unverified users, we observe the existence of three unverified accounts among the top 20 users who shared tweets with warning labels (recall, from the dataset section, that we collect the tweets that verified accounts retweeted or quoted by unverified accounts). This indicates that Twitter's moderation mechanism is not only limited to verified users. Finally, we note that 6 out of the top 20 users with tweets that had warning labels were suspended by Twitter (as of January 9, 2021). This highlights that the continuous dissemination of questionable content that leads to the addition of warning labels is likely to result in hard moderation interventions (i.e., user suspensions).

Take-Aways. The main take-away points from our analysis on warning labels and Twitter users are:
1. Most of the warning labels on Twitter, between November 2020 and December 2020, were related to the 2020 US elections. Also, we find different temporal patterns in the use of warning labels, with a few of them being short-lived (less than a week) and some of them being long-
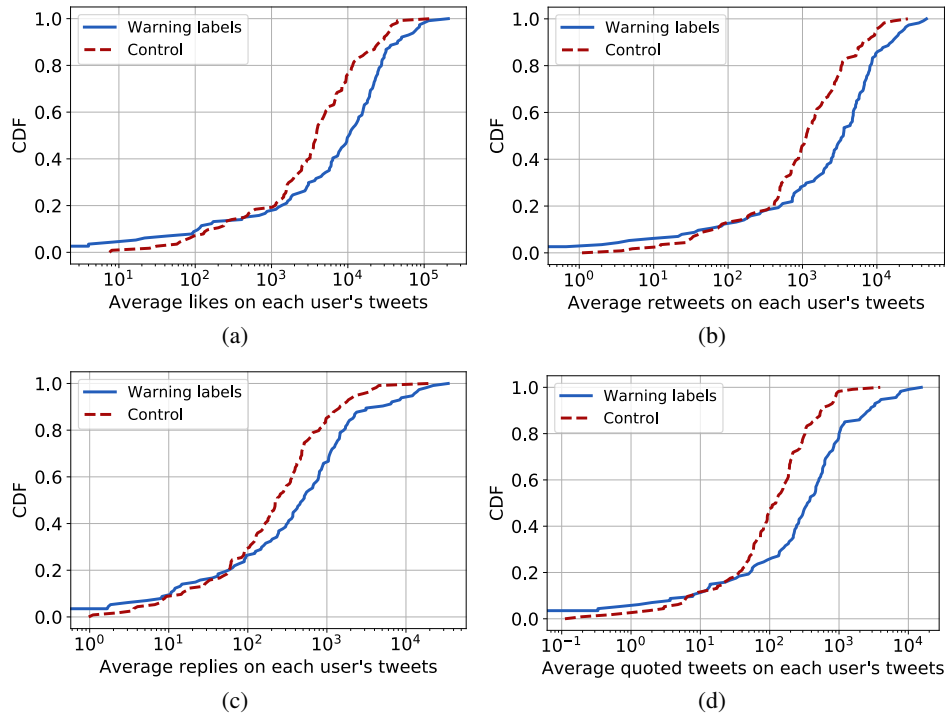
(a)

(b)

(c)

(d)

Figure 4: Mean engagement metric for each user for tweets with warning labels and for tweets without warning labels.
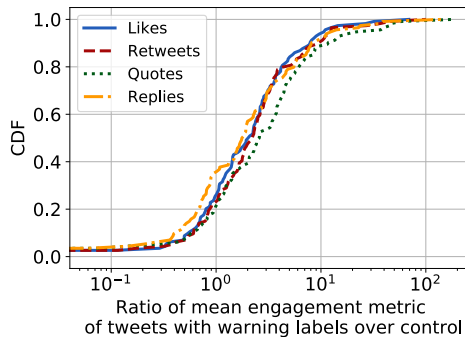

Figure 5: CDF of the ratio of the mean engagement metric for tweets with warning labels over control.

lived (across several months).

2. We find warning labels used to inform users about manipulated multimedia, while some warning labels are in languages other than English (i.e., Portuguese). This highlights the efforts put in soft moderation interventions and some of the challenges (e.g., tracking claims across multiple information formats or languages).

3. The majority of tweets with warning labels (72%) are shared by Republicans, while Democrats are more likely to comment on tweets with warning labels using Twitter's quoting functionality (56% of the tweets compared to 16% for Republicans). These results likely indicate that Republicans, during the 2020 US elections, shared more questionable content that was eventually flagged by Twit-

ter moderators.

4. The continuous dissemination of potentially harmful information that is annotated with warning labels can lead to hard moderation interventions like permanent user suspensions. We find that 6 out of the 20 top users, in terms of sharing tweets with warning labels, were permanently suspended by Twitter as of January 9, 2021.

RQ2: Engagement Analysis

The goal of warning labels is to provide adequate information on tweets that include questionable content and might be harmful to users or society. Thus, we expect that users who see content that is annotated with warning labels are likely to cause them to be less willing to engage with or re-share such content (Mena 2020). In this section, we aim to quantify the differences in the engagement between tweets that include warning labels and tweets that do not. Our empirical analysis can quantify how effective are the warning labels on Twitter through the lens of user engagement.

For each user in our dataset, we extract two sets of tweets: 1) tweets that have warning labels; 2) a control dataset of tweets that do not have warning labels. Note that we limit our analysis to the 115 users with at least three tweets with warning labels to make sure that our analysis is not influenced by one or two tweets. Then for each engagement metric in our dataset, we calculate the mean number that each group of tweets (warning label tweets and control) had for each user. Our analysis takes into account four engagement metrics: 1) *Likes* (how many times the tweet was liked by other users); 2) *Retweets* (how many times the tweet was
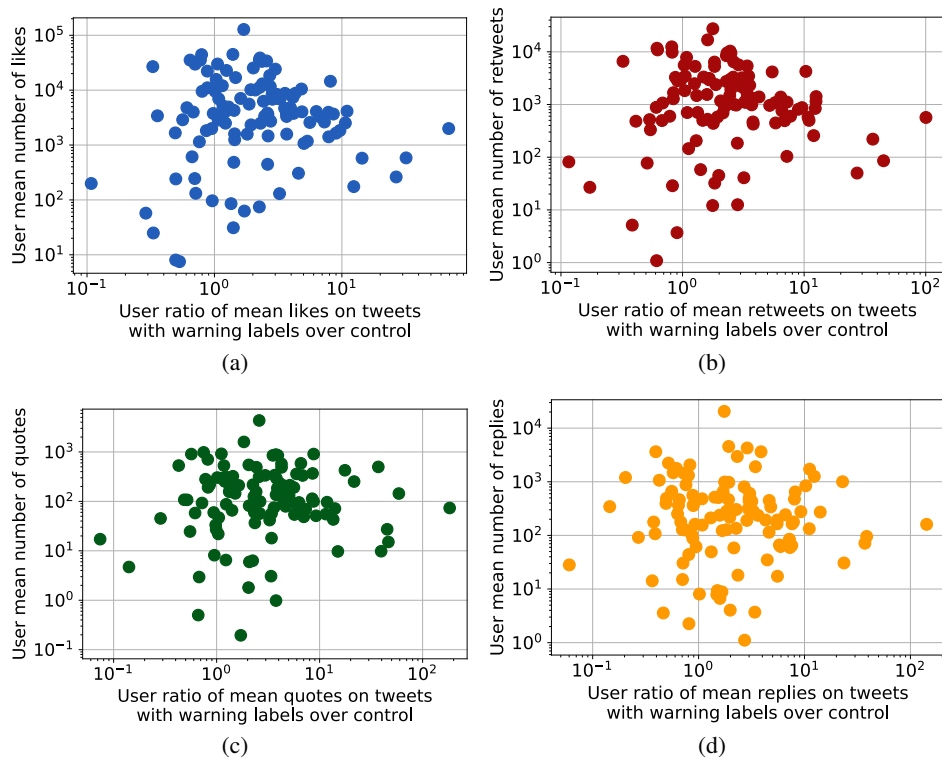
Figure 6: Number of followers and the ratio of the mean engagement for tweets with warning labels over the control dataset.

retweeted by other users); 3) *Quotes* (number of other tweets that retweeted the tweet with a comment); and 4) *Replies* (number of replies that the tweet received).

Fig. 4 shows the Cumulative Distribution Function (CDF) of the average number of likes/retweets/quotes/replies of tweets with and without warning labels per user. For each engagement metric, we perform two-sample Kolmogorov-Smirnov statistical significance tests, finding that in all cases, the engagement of tweets with warning labels is significantly different compared to tweets without warning labels ($p < 0.01$) We observe that, for all four engagement metrics, users receive increased engagement on tweets that have warning labels.

For likes (see Fig. 4(a)), we find a median value of 10,303.9 average likes per user for tweets with warning labels, whereas, for the control dataset, we find a median value of 3,834.3 (2.6x less than warning labels). For retweets (see Fig. 4(b)), we find a median value of 3,533 average retweets per user for tweets with warning labels, while for the control dataset, the median value is only 1,129.2 (3.1x decrease compared to the warning labels). For replies (see Fig. 4(c)), we find a median value of 235.7 replies for the control dataset, while for warning labels, the median value increases to 494 (2.1x increase over the control dataset). For quotes (see Fig. 4(d)), we find a median value of 350.6 average quotes per user for the warning labels datasets, whereas, for the control dataset, we find a median value of 122.9 quotes (2.8x decrease compared to warning labels). Also, from Fig. 4, we can observe a small proportion of users

who have less engagement on the warning labels dataset. To quantify the proportion of users who have more engagement on control tweets over the tweets that had warning labels, we plot the ratio of the mean number of each engagement metric on tweets with warning labels over the control dataset (see Fig. 5). When this ratio is below 1, user's control dataset had more engagement compared to the user's warning labels dataset. We find that 26%, 23%, 21%, 35% of the users had more engagement on their control tweets over the ones with warning labels for likes, retweets, quotes, and replies, respectively.

From our analysis thus far, it is unclear which set of users have increased vs. decreased engagement on tweets with warning labels over the control dataset. To assess whether there is a correlation between the overall engagement that user receives and whether a user will receive increased or decreased engagement on tweets with warning labels, we plot the overall engagement (i.e., mean engagement metric for all the user's tweets) and the ratio of engagement on warning labels over the control dataset (see Fig. 6). We observe that for all engagement metrics, most of the users that have on average high engagement on their content (i.e., over 1K likes, over 100 retweets, over 100 quotes, and over 100 replies) also receive an increased engagement on tweets with warning labels over the control (note that the ratio for these users is in most of the times between 1 and 10).

Take-Aways. The main points from this section are:
1. Tweets with warning labels had more engagement compared to tweets without warning labels. This is likely be-

cause Twitter prioritizes content that receives high engagement in their system (Roth and Pickles 2020).

2. We find that 65%-79% (depending on engagement metric) of the users receive increased engagement on their tweets with warning labels compared to tweets without warning labels.

3. By looking at the users that have increased vs. decreased engagement on tweets with warning labels compared to the control dataset, we find that most users who have high engagement in general have also increased engagement on tweets with warning labels.

RQ3: User Interaction with Tweets and Warning Labels

In this section, we study how users interact with tweets that have warning labels. To do this, we use Twitter's quote functionality, where users can retweet a tweet with a comment. Specifically, we qualitatively analyze three sets of tweets; 1) the 50 tweets that quote other tweets and Twitter includes warning labels on **both** tweets; 2) 122 tweets (out of the 169) that quote other tweets and Twitter includes a warning label **only** on the top tweet (i.e., user's comment). The 47 other tweets had a quoted tweet that was deleted when we tried to assess them qualitatively; 3) 150 randomly selected tweets that quote another tweet that includes a warning label. We qualitatively analyze all three sets of tweets to understand how users interact with people that share content that is annotated with warning labels, how users interact with questionable content (e.g., false claims), and how users discuss or perceive the existence of warning labels on Twitter. We employ qualitative thematic analysis (Braun and Clarke 2006) to identify a set of codes from our dataset, code the tweets in our samples, and then group the tweets into themes. Below, we present the main themes for each sample.

Quoted Tweet and Comment Include Warning Labels

Intuitively, when both the quoted tweet and the comment tweet above include warning labels (e.g., Fig. 7), one expects that both tweets include information that is questionable or potentially harmful. Here, we qualitatively analyze the tweets in our dataset to verify if this is the case and what are other cases where both the quoted and the comment tweet include warning labels.

Reinforcing False Claims. The majority of the comments above the quoted tweets aim to retweet and reinforce the false claim included in the quoted tweet (86%, 43 out of the 50). Two of them achieve this using a single word ("this" or "true"), two of them use videos, five of them achieve it by tweeting a single hashtag (#stopthesteal and #ExposeDominion that both refer to election fraud claims), while the rest of the comment tweets use text to reinforce the claim. The fact that some of the comments comprise only a single word shows that adding warning labels to tweets requires considering the context and other quoted tweets and not focusing on the tweet in isolation. Also, in 2 out of the 43 comments that reinforce the claim of the quoted tweet, the users share their anti-censorship opinions or disputing the fact that
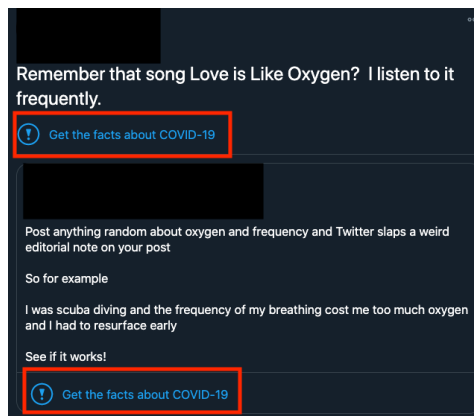

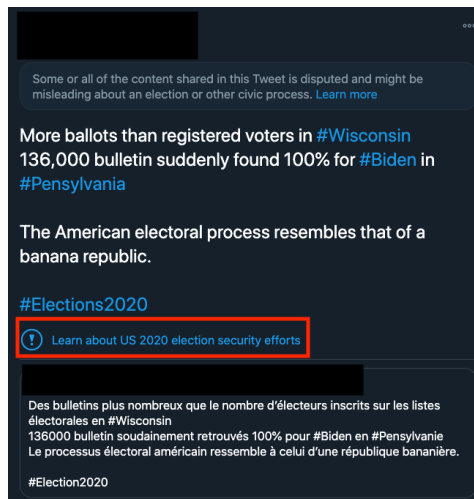Figure 7: Example of a wrong addition of warning labels.

Figure 8: A quoted tweet that is not flagged likely because is in French.

the content should be labeled (i.e., *"Say NO to Big Tech censorship!"* and *"Twitter labeled this tweet as disputed.... What exactly is Twitter disputing here?"*). These results confirm the findings from (Saltz, Leibowicz, and Wardle 2020).

Testing Warning Labels. We find one tweet where the user commented with the same content as the quoted tweet, likely to verify if his comment will eventually get a warning label.

Incorrect Warning Labels. We find one specific case where the warning labels were seemingly incorrectly put (see Fig. 7). Both the comment and the quoted tweet had the warning label "Get the facts about COVID-19" and both were including the terms oxygen and frequency/frequently. This likely indicates that Twitter employs automated means to attach warning labels and in some cases, warning labels are incorrectly added to some content.

Warning Label on Comment Above

Next, we investigate cases where users quote a tweet with no warning label, and subsequently, their comment tweet receives a warning label (e.g., Fig. 8).
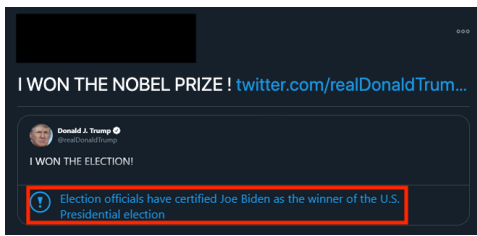
Figure 9: Example of a tweet that is mocking the author or the content of the quoted tweet.

Making False Claims. We find 45 tweets (36%) that comment on real-world events, news, or facts about the election, and make false claims about the election (e.g., claims about election fraud).

Reinforcing Questionable Content. In 18 tweets (14%), the comment above reinforces questionable content included in the quoted tweet and makes a claim even more questionable or harmful, hence getting flagged by Twitter.

Inconsistencies on Warning Labels. We find several cases where there are inconsistencies with the inclusion of warning labels. Specifically, we find 28 cases (23%) where both the quoted tweet and the comment hint at election fraud during the 2020 US elections, yet only the quoted tweet includes a warning label. In 7 of these cases, the comment makes a similar claim with the quoted tweet with the difference that it uses a video instead of text. This highlights the challenges in flagging content on social media platforms and, in particular, flagging the same information across multiple diverse formats (i.e., text, images, videos). Also, we find another case with inconsistencies related to the use of language. In this case, the quoted tweet and the comment above share the same information but in different languages (quoted tweet in French and comment above in English), yet only the English comment includes a warning label (see Fig. 8).

Updates on Warning Labels. During our qualitative analysis, we observed that Twitter occasionally updates the warning labels on some tweets. In particular, we find many instances where Twitter changed the warning label from "Multiple sources called this election differently" to "Election officials have certified Joe Biden as the winner of the U.S. Presidential election". This highlights that Twitter continuously refines warning labels and that it is likely that they update warning labels on content to make the warning label clearer or stronger.

Warning Label on the Quoted Tweet

Here, we aim to understand how users interact with content that includes warning labels by looking into tweets that quote content with warning labels (e.g., Fig. 9). We find various behaviors ranging from mocking the author/content of the quoted tweet, debunking false claims on the quoted tweet, reinforcing the false claims, and sharing opinions on Twitter's warning labels. We provide more details below.

Mocking or Sharing Emotions about the Author/Content of the False Claim. We find 37 tweets that mock the content or the author of the tweet that includes a warning label. For instance, when President Trump tweeted the tweet in Fig. 1, several users quoted that tweet and made absurd claims about themselves like *"I WON THE NOBEL PRIZE !"* (see Fig. 9) and *"Let me try... I AM BEYONCE!!"*. Other users quoted tweets with warning labels to express their emotions on the content or the author of the tweet: 4 tweets calling the quoted tweet author a liar, 4 tweets calling the author a loser, 6 tweets expressing their disgrace for the content of the tweet, and 1 tweet expressing embarrassment.

Debunking False Claims. We find 19 tweets that debunk false claims that are in quoted tweets. For instance, a user quoted a tweet shared by President Trump and wrote: *"President Trump just tweeted again about claims of "secretly dumped ballots" for Biden in Michigan. This is false. These claims are based on screenshots of a mistaken unofficial tally on one site's election map that was caused by a typo that was corrected in about 30 minutes."*

Reinforcing False Claims. Similarly to the tweets where both the quoted and the comment above had warning labels, we find 6 tweets that were reinforcing false claims that exist on the quoted tweets.

Sharing Opinion on Warning Labels. We find 6 tweets that share users' opinions on warning labels and how effective they are. Specifically, one tweet indicates that the quoted tweet includes a warning label and two tweets question how effective the warning labels are and request stronger and more straightforward labels. Also, we find three tweets that call for hard moderation interventions (i.e., user bans), in particular asking Jack Dorsey (Twitter's Chief Executive Officer) or Twitter Support to ban the account of President Trump due to the spread of false claims (e.g., *".@jack @Twitter make this lying stop! Your warnings of him lying just are not enough. #BanTrump"*). Interestingly, we find one tweet where the comment reinforces the false claim included in the quoted comment by claiming that Twitter tries to cover up the election fraud using warning labels.

Take-Aways. The main take-away points from our qualitative analysis are:

1. We find various user interactions with tweets that have warning labels, such as debunking false claims, mocking users that tweeted questionable content, or reinforcing false claims despite the inclusion of warning labels.
2. Soft moderation intervention systems are not always consistent, as we find several cases where content should have warning labels but it does not. E.g., we find cases where videos share the same information with textual tweets that include warning labels, however, the tweet with the video does not include a warning. Another example is with content across various languages. These cases show the challenges that exist on large-scale soft moderation systems.
3. We find a case where the warning label was incorrectly added, likely due to automated means. This shows the need to devise systems that rely on human moderators that get signals from automated means (i.e., the human makes the final decision), hence decreasing the likelihood of such cases.

Discussion & Conclusion

In this work, we performed one of the first characterizations, based on empirical data, of soft moderation interventions on Twitter. Using a mixed-methods approach, we analyzed the warning labels, the users that share tweets that have warning labels, and the engagement that this content receives. Also, we investigated how users interact with such content and what are the challenges and some inconsistencies that exist on large-scale soft moderation systems.

Our user analysis showed that 72% of the tweets with warning labels were shared by Republicans, which likely indicates that Republicans were sharing more questionable/harmful content during the 2020 US elections. Overall, this finding prompts the need for greater transparency by social media platforms to ease concerns related to censorship and possible moderation biases towards a specific political party (Clayton 2020).

Our engagement analysis showed that tweets with warning labels had more engagement than tweets without warning labels. This finding is in contrast with previous work by (Mena 2020) that found that users were less willing to share content with warning labels. This likely indicates that in political discussions on Twitter, users are affected more by their own political ideology/biases rather than warning labels added by Twitter. Another possible explanation is that the tweets that eventually received warning labels were potentially harmful and were getting a large engagement even before the addition of the warning label. Nevertheless, the fact that tweets with warning labels receive a substantial engagement by users, highlights the need for stricter soft moderation designs. Indeed, Twitter started imposing further restrictions on tweets like the ability of Twitter users to retweet, like, or reply to tweets that include content that can be extremely harmful for the platform and society (e.g., may cause real-world violence) (Gadde and Beykpour 2020).

Our qualitative analysis showed that users further debunk false claims using Twitter's quoting mechanism, they mock the user/content of the tweet with a warning label, and they reinforce false claims (despite the warning labels). Similarly to (Saltz, Leibowicz, and Wardle 2020), we observed empirical evidence of users perceiving warning labels as an act of censorship by the platforms and publicly disseminating their opinion on Twitter. Also, we found some inconsistencies in content that should be flagged across multiple information formats or languages. This highlights some of the challenges that exist when designing moderation systems that are applied to enormous and diverse platforms like Twitter. Taken altogether, as a research community we need to further study such moderation systems to fully understand how they work and what their caveats are to increase their effectiveness, consistency, fairness, and transparency.

Limitations. Our work has some limitations. First, we analyzed mainly politics-related content, shared during a short period of time (two months), on a single platform (Twitter). Thus, it is unclear whether our results hold in contexts not related to politics or to soft moderation systems that exist on other platforms like Facebook (as it has different platform affordances and design of soft moderation interventions).

Also, our engagement analysis does not account for the content of tweets, hence we do not investigate whether the increased engagement on tweets with warning labels is due to the dissemination of more controversial or sensationalistic content that is likely to attract more users. Finally, since we do not know exactly when a soft moderation intervention happened and how the engagement changed over time, we do not analyze whether the warning labels were added because the tweets received large engagement in advance.

Validation of the Political Bias Inference API

(Kulshrestha et al. 2017) measured the performance of the Political Bias Inference API; they found an accuracy of 92% for US senators in 2017 and 85% for Twitter users that self-identified their political views on Twitter. Given that our dataset also includes users that are not US senators and that do not self-identify their political views on Twitter, we evaluated the performance of the Political Bias Inference API in our dataset. To validate the performance of the Political Bias Inference API, we extracted a sample of 200 users that were independently classified as Republican/Democrat/Neutral by the first author of this study. The sample includes the top 100 users in our dataset (in terms of the tweets with warning labels/quoted tweets) and another 100 randomly sampled users (we ensured no overlap between the two sets). The manual classification task for each user was undertaken as follows. First, we inspected the Twitter user profile to check if they report their political views/party. Second, we searched on Google to find more about the user and find references mentioning their political views. Finally, we searched for tweets posted by the user in consideration that discussed politics to infer their political views. Then, we compared the manual classification with the inferences from the Political Bias Inference API; we found an average accuracy of 80.5% (82% for the top 100 users and 79% of the randomly selected 100 users). Even though the API has a considerable number of misclassifications, we believe that it can provide a meaningful signal with regards to the political views of the users in our dataset.

References

Bode, L.; and Vraga, E. K. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication* 65(4): 619–638.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3(2): 77–101.

Chandrasekharan, E.; Jhaver, S.; Bruckman, A.; and Gilbert, E. 2020. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *arXiv preprint arXiv:2009.11483* .

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *ACM on Human-Computer Interaction* 1(CSCW): 1–22.

Clayton, J. 2020. Social media: Is it really biased against US Republicans? https://bbc.in/3sYzCc5. Accessed: 2021-04-06.

Evelyn, K. 2021. Capitol attack: the five people who died. https://www.theguardian.com/us-news/2021/jan/08/capitol-attack-police-officer-five-deaths. Accessed: 2021-04-06.

Gadde, V.; and Beykpour, K. 2020. Additional steps we're taking ahead of the 2020 US Election. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html. Accessed: 2021-04-06.

Geeng, C.; Francisco, T.; West, J.; and Roesner, F. 2020. Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact. In *arXiv preprint arXiv:2012.11055*.

Ghosh, S.; Sharma, N.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*, 575–590.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Guynn, J. 2016. "Massive rise" in hate speech on Twitter during presidential election. https://eu.usatoday.com/story/tech/news/2016/10/21/massive-rise-in-hate-speech-twitter-during-presidential-election-donald-trump/92486210/. Accessed: 2021-04-06.

Hutchinson, A. 2020. Twitter Updates Hate Speech Policy to Include Links to "Hateful" Content. https://www.socialmediatoday.com/news/twitter-updates-hate-speech-policy-to-include-links-to-hateful-content/582482/. Accessed: 2021-04-06.

Kaiser, B.; Wei, J.; Lucherini, E.; Lee, K.; Matias, J. N.; and Mayer, J. 2020. Adapting Security Warnings to Counter Misinformation. *arXiv preprint arXiv:2008.10772* .

Kulshrestha, J.; Eslami, M.; Messias, J.; Zafar, M. B.; Ghosh, S.; Gummadi, K. P.; and Karahalios, K. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *CSCW*, 417–432.

Mena, P. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet* 12(2): 165–183.

Merrer, E. L.; Morgan, B.; and Trédan, G. 2020. Setting the Record Straighter on Shadow Banning. *arXiv preprint arXiv:2012.05101* .

Messias, J. 2017. Political Bias Inference API. https://github.com/johnnatan-messias/bias_inference_api. Accessed: 2021-04-06.

Moravec, P. L.; Kim, A.; and Dennis, A. R. 2020. Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. *Information Systems Research* 31(3): 987–1006.

Myers West, S. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20(11): 4366–4383.

Newell, E.; Jurgens, D.; Saleem, H. M.; Vala, H.; Sassine, J.; Armstrong, C.; and Ruths, D. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *ICWSM*, 279–288.

Nyhan, B.; and Reifler, J. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32(2): 303–330.

Pennycook, G.; Bear, A.; Collins, E. T.; and Rand, D. G. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66(11): 4944–4957.

Pennycook, G.; Cannon, T. D.; and Rand, D. G. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147(12): 1865.

Pushshift. 2020. Dataset of Verified Users. https://files.pushshift.io/twitter/TW_verified_users.ndjson.zst. Accessed: 2021-04-06.

Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and West, R. 2020. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *arXiv preprint arXiv:2010.10397* .

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research* 3.

Rosen, G. 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/. Accessed: 2021-04-06.

Roth, Y.; and Pickles, N. 2020. Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html. Accessed: 2021-04-06.

Saleem, H. M.; and Ruths, D. 2018. The Aftermath of Disbanding an Online Hateful Community. *arXiv preprint arXiv:1804.07354* .

Saltz, E.; Leibowicz, C.; and Wardle, C. 2020. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. *arXiv preprint arXiv:2011.12758* .

Seo, H.; Xiong, A.; and Lee, D. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *WebSci*, 265–274.

Sharma, N. K.; Ghosh, S.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Inferring who-is-who in the Twitter social network. *ACM SIGCOMM CCR* 42(4): 533–538.

Solon, O. 2016. Facebook's failure: did fake news and polarized politics get Trump elected? https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories. Accessed: 2021-04-06.

Swire-Thompson, B.; DeGutis, J.; and Lazer, D. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition* .

Twitter. 2020. Notices on Twitter and what they mean. https://help.twitter.com/en/rules-and-policies/notices-on-twitter. Accessed: 2021-04-06.

Twitter Inc. 2021. Permanent suspension of realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension.html. Accessed: 2021-04-06.

Wood, T.; and Porter, E. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41(1): 135–163.

Zannettou, S. 2021. Dataset of Soft Moderation Interventions on Twitter. https://github.com/zsavvas/Soft-Moderation-Interventions-Twitter. Accessed: 2021-04-06.