IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 June 2023, accepted 8 July 2023, date of publication 24 July 2023, date of current version 18 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298557

RESEARCH ARTICLE

Of Stances, Themes, and Anomalies in COVID-19 Mask-Wearing Tweets

JWEN FAI LOW[1], BENJAMIN C. M. FUNG[1], (Senior Member, IEEE), AND FARKHUND IQBAL[2], (Member, IEEE)
[1]School of Information Studies, McGill University, Montreal, QC H3A 1X1, Canada
[2]College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

Corresponding author: Benjamin C. M. Fung (ben.fung@mcgill.ca)

The work of Jwen Fai Low and Benjamin C. M. Fung was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grant RGPIN-2018-03872 and the Canada Research Chairs Program under Grant CRC-2019-00041. The work of Farkhund Iqbal was supported by Zayed University, United Arab Emirates, through the Cluster Project under Grant R16083 and the Provost Research Fellowship Grant R20093.

ABSTRACT COVID-19 is an opportunity to study public acceptance of a ''new'' healthcare intervention, universal masking, which unlike vaccination, is mostly alien to the Anglosphere public despite being practiced in ages past. Using a collection of over two million tweets, we studied the ways in which proponents and opponents of masking vied for influence as well as the themes driving the discourse. Pro-mask tweets encouraging others to mask up dominated Twitter early in the pandemic though its continued dominance has been eroded by anti-mask tweets criticizing others for their masking behavior. Engagement, represented by the counts of likes, retweets, and replies, and controversiality and disagreeableness, represented by ratios of the aforementioned counts, favored pro-mask tweets initially but with anti-mask tweets slowly gaining ground. Additional analysis raised the possibility of the platform owners suppressing certain parts of the mask-wearing discussion.

INDEX TERMS Social media, Twitter, censorship, information diffusion, ratiometrics, stance classification, theme classification, summarization, machine learning, transformers.

I. INTRODUCTION

Prior infections and vaccinations have proven insufficient in combating COVID-19.[1] Prior infections do not fully protect against reinfections [3], [4]. Vaccine effectiveness is reduced against new SARS-CoV-2 variants such as Delta [5], [6], [7] and Omicron [8], [9], [10], [11], necessitating bivalent, trivalent, and tetravalent vaccines [12], [13], [14], [15]. Vaccination campaigns are stymied by vaccine hesitancy [16], [17], [18] and insufficient global supplies [19], [20], [21]. To influence the public to accept protective measures to combat the COVID-19 pandemic, capturing hearts and minds on social media [22], [23], [24] is essential due to the declining trust in mainstream media [25] and the growing share of social media users, which consists of 55.1% of the global population [26] and seven-in-ten Americans [27].

Most research focused on the vaccine discourse on social media; we opt to study the equally crucial conversations on masking to chart the evolution of the position adopted and the issues raised by social media users. Masks are an important non-pharmaceutical intervention (NPI) in protecting against infection, with some healthcare experts saying that masks are comparable or better than vaccines at stopping SARS-CoV-2 [28], [29], [30]. Experts also compared masking against SARS-CoV-2 to wearing condoms against the endemic HIV threat [31], [32].

Our study assigned stances and themes to a large corpus of tweets using a DistilBERT classifier trained on hand-labeled examples. Stance and theme information are then used in conjunction with the counts of likes, retweets, and replies as well as the ratios of the counts to study temporal trends. Further probing was done by examining TF-IDF

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

[1] A disease does not disappear after transitioning from pandemic to endemic [1], [2]. Malaria, HIV, tuberculosis, etc. remain endemic threats.

word importance and summaries generated by BART. Our analysis showed that the largest share of tweets are those with a stance supportive of universal masking and featuring a theme of encouraging mask-wearing at the start of the pandemic. As the pandemic progressed, tweets opposing universal masking and criticizing others' masking behavior gained ground at the expense of supportive tweets, although they never became the majority. The decline in popularity, non-controversiality, and agreeableness of supportive tweets abruptly and briefly reversed their trend during an anomalous event between late May and early June. Closer examination of this disruptive anomalous event and cross-referencing with information from another study on the masking conversations on Twitter strongly indicate an attempt to censor tweets of particular themes and stance during this period. Other notable findings include (1) user attention concentrated on a small number of conversations being more effective in altering overall stance than dispersed attention (Section IV-D3), (2) explicitly politicized debates garnering more attention than non-politicized ones (Section IV-D6), and (3) narrative from tweets being capable of supplanting narrative of other tweets with the same theme but opposing stance (Section IV-D3). Our investigation into the anomalous event suggests that future studies involving social media data should not presume platform neutrality and actively work to account for content moderation from platform owners altering the observed data distribution.

II. BACKGROUND

A. SOCIAL MEDIA DISCOURSE ON COVID-19

COVID-19 struck the global population in an age of widespread social media usage. Researchers leveraged the unprecedented access to social media data [33] to study thoughts and concerns voiced by social media users, e.g., [34], and how information from social media influenced their thought processes, e.g., [35]. Massive curated and processed datasets, e.g., [36], were publicly released to facilitate investigations. Of interest to many researchers is the "infodemic" occurring in parallel with the real-life pandemic, where the indiscriminate sharing of dubious information or outright misinformation risks drowning out sound public health advice. Most researchers studied misinformation in general [37], [38], [39], [40] while some focused on vaccine hesitancy misinformation [41] or on profiling users who share information on controversial treatments such as hydroxychloroquine [42]. Researchers also inferred the sentiments expressed within the text of social media conversations [43], [44], with [45] taking the unique route of interviewing users about their sentiment *towards* information on COVID-19 found on social media. Researchers have also looked into debates (explicitly oppositional unlike conversations) occurring on social media, such as those concerning vaccination [46]. While stances are touched upon in [46], they are not studied in-depth, with the authors themselves remarking that "[d]eeper analyses on text content should be performed to identify users who are either criticizing the anti-vax movements or supporting them". Stances on issues related to COVID-19 have received comparatively little research attention overall.

B. SOCIAL MEDIA DISCOURSE ON MASK-WEARING

There is only one other study on mask-wearing stance in the context of COVID-19 using social media data [47]. In this study by Cotfas et al., they found that the best stance classification performances were obtained from large language models, BERT and Roberta, which are similar to the DistilBERT we used to classify both stance and themes (Section III-B5). However, their study differed from ours in some major areas: the data collection process, the time period covered, the absence of themes, and the absence of engagement metrics and ratiometrics analyses. The tweets they collected all contain keywords referring to masking *and* COVID-19. In contrast, we collected all tweets with just masking keywords and used our DistilBERT classifier to remove tweets unrelated to COVID-19, which is sufficiently discriminative (Section IV-C) to produce a clean dataset. They covered all twelve months of 2020 whereas we covered the initial six months, which we further reduced to the March and June period to improve the accuracy our engagement metric and ratiometrics analyses. Their cleaned dataset exhibited a noisy pattern in stance proportions in January, demonstrating the issue of including earlier periods in analyses. Without classifying themes, Cotfas et al. [47] would not have been able form a cohesive picture of the drivers behind online discussions, e.g., the decline of tweets promoting/discouraging mask-wearing and the late emergence of tweets framing mask-wearing as individualistic/collectivistic behavior (Figure 5). Without examining the engagement metrics of likes, retweets, and replies, Cotfas et al. [47] would have been able to gauge audience receptiveness to the stances and themes publicly expressed by social media users.

The study by Sanders et al. [48] also examined COVID-19 mask-wearing conversations on Twitter, but they focused on sentiments instead of stances. Sentiments are emotive. Positive (e.g., joy) and negative (e.g., anger) sentiments on tweets are not equal to supportive and oppositional stances on issues [49]. In [48], sentiments were labeled with lexicon-based VADER and topics were discovered via clustering of embedded representations of tweet texts instead of manual coding. Consequently, the topics are only mostly reflective of the proximity of words in the embedding space and are decontextualized from the subject of interest, the COVID-19 mask-wearing conversation. This can be observed in [48] having seemingly overlapping topics, e.g., Cluster 8 with keywords of `wearamask / maskwearing / masking`, Cluster 14 with `face / facemask / coronavirus`, and Cluster 13 with `coronavirus / face / make`. Furthermore, the researchers showed no intention of associating tweets with more than one topic. In our research, a tweet can be composed of multiple themes.

Although there exist studies of an earlier major respiratory disease outbreak — the H1N1 pandemic — using social media data, masks were not their concern. A 2019 thematic analysis of H1N1-related tweets [50] contained no mention of masks, demonstrating the unimportance of masks in conversations on pandemics until COVID-19. A 2010 study [51] has a single finding related to masks, which was that a campaign comparing masks for H1N1 with condoms for AIDS downplayed H1N1's risks. A 2011 study focusing on US public concern over H1N1 [52] similarly neglected masks, only remarking that spikes in the number of mask-related tweets seemed to coincide with the CDC's messages and exhibited a downward temporal trend.

As for even earlier respiratory disease outbreaks such as the 2004 SARS outbreak, the interaction between them and social media have not been studied, likely because these earlier outbreaks were not truly global and social media platforms have not gained prominence (Facebook and Twitter opened to the public around 2006). For seasonal flu endemic to temperate climate zones, they have not generated much social media activity on masking as a topic likely due to the absence of risk messaging in mass media [53] and also the lack of encouragement for mask-wearing from public health authorities.

III. METHODS

A. DATA

To collect tweets discussing mask-wearing, we used *Twint*, a scraper that bypasses Twitter API's rate limits by using Twitter's search operators. Twint has been used in academic research dealing in a broad range of topics including COVID-19 discourse [54], [55], disinformation campaigns [56], and citizen engagement [57]. Twint collects only publicly viewable tweets and is incapable of accessing protected tweets, which are only visible to the followers of the protected tweets. For the data collection's starting point, we selected January 1, 2020, the date on which China alerted WHO of pneumonia cases of indeterminate cause in Wuhan and the knowledge of a respiratory disease outbreak began entering public consciousness in the Anglosphere.

The search keywords used are "wear" and "mask". These keywords were chosen after performing extensive exploratory searches that demonstrated the non-viability of other keywords. "Face", "put", "on", "don", "filter", "cover", "covering", "facemask", "faceshield", "facepiece", "respirator", "scarf", "bandanna", "cloth", "elastomeric", "N95", "KN95", "FFP", "FFR", "protection", "shield", and their various combinations were excluded either due to their non-specificity or their relative unpopularity. Only "wear" and "mask" proved sufficiently unambiguous and popular to retrieve tweets where the vast majority are about donning protective face covering against COVID-19.

The search engine is case-insensitive and even includes tweets with #wearmask hashtag in its results. However, different forms of the keywords "wear" and "mask", e.g., "wearing", "worn", "masked", "masks", were not retrieved by Twitter's search engine. Search results included tweets which have neither "wear" nor "mask" in them; the keywords were instead found in the tweets' authors' names, as it was in vogue for Twitter users to include variations of "wear a mask" in their display names. We excluded these tweets without keywords. The final dataset consists of 2,100,932 tweets from January 1, 2020 to June 21, 2020. We considered refining the results by requiring the tweets to include a term associated with COVID-19 such as "coronavirus", "pandemic", or #stayhome, but we decided against it as our exploratory searches found that most tweets did not mention COVID-19 and assumed that the context in which mask-wearing is being discussed would be self-evident, e.g., "Pence refused to wear a mask" is far more common than "Pence did not wear a mask, which is in violation of coronavirus health regulations". Our decision is validated by our analysis showing many of the retrieved tweets to be COVID-19 related (Section IV-C).

B. CLASSIFICATION

We first manually labeled a subset of the collected tweets, then trained a machine learning classifier on this human-labeled set, and finally labeled the remaining unlabeled entries using the classifier. An earlier paper studying vaccination stances of Twitter users [58] used a similar labeling strategy; they hand-labeled 2% of tweets and machine-labeled the rest. While the number of human-labeled training examples may appear insufficient, this is compensated by our reliance on DistilBERT (Section III-B5), a transformer-based architecture that incorporates knowledge acquired from pre-training on very large text corpora. Fine-tuning a pre-trained DistilBERT for other tasks requires comparatively much smaller datasets. Other transformer-derived classifiers, such as BERT and RoBERTa, have achieved state-of-the-art results for domains such as biomedical and legal texts in zero-shot and few-shot large (LMTC) and extreme multi-label text classification (XMTC) tasks, which have inherent severe class imbalance issues, without necessarily resorting to data augmentation.

Each tweet has only one stance. Stances are mutually exclusive, and the stance assigned to a tweet is representative of the entire tweet. A tweet, however, can have one or more themes. Themes can encompass the entire tweet or sections of it, and the same sections of text can have overlapping themes. Stance and themes are separate attributes. Although we will explore potential linkages between these attributes (Section IV-B), a tweet's themes should not be coupled with its stance from the outset.

1) STANCE STANDARD

For every perspective (tweet), we test it against the following claim (standard)[2] to determine the tweet's stance: *"everyone wearing masks in public will help end the COVID-19*

[2] The "perspective" and "claim" terminology is from [59].

pandemic". "Masks" include makeshift fabric face coverings such as bandannas and scarves to professionally manufactured filtering facepiece respirators but excludes costume mask with visible gaps in the mouth and nose areas. "Public" covers all areas outside a person's place of residence without any exception, be it engaging in vigorous activities, the presence of social distancing, or being outdoors. "Everyone" is literal and does not brook exceptions either, not even people suffering from mental health issues such as anxiety attacks and PTSD or health/respiratory issues such as asthma and COPD.

The standard that we test against may appear lenient in what passes for masks and excessive in what constitutes "public" and "everyone", but the wording of this standard is an attempt to harmonize technical terms and specific interests of healthcare experts with the non-specialized terms and broad concerns found among Twitter users.

Our standard is based upon a survey of peer-reviewed scientific literature on mask usage against respiratory diseases. The prevailing consensus among the systematic reviews and meta-analyses [60], [61], [62], [63], [64] is that masks could be beneficial in limiting the transmission of respiratory viral infections. Evidence from ecological studies [64], [65], [66] also favors community-wide mask-wearing as an effective measure in controlling the spread of COVID-19. There is a lack of agreement on how different *types* of masks perform as source control and protection in community settings [60], [61], [62], [63], [64], [65], [67], [68], [69], hence the broad definition used in the standard.

We consciously avoided basing the standard upon advisories issued by public authorities, both health (e.g., WHO and CDC) and non-health, as they have changed over time and, in the case of non-health authorities, they may not be based entirely on sound evidence and may be influenced by practical (e.g., material supply) and/or political considerations. The WHO issued an interim guidance on April 6, 2020 [70] recommending against mask usage in community settings except for those who are symptomatic and people in contact with them, but on June 5, 2020 [71], this was changed to encourage more widespread mask usage among healthy individuals. Cities and counties in California and Colorado have banned valved facemasks out of fear that they function poorly at source control [72], but in December 2020 the CDC published a report showing that masks with exhalation valves are no worse at source control than surgical and cloth masks [73].

2) STANCE LABELING PROCEDURE

There are four stance labels: negative (oppose), neutral, positive (support), and out-of-topic. Their corresponding numeric codes are 1, 3, 5, and 9. Out-of-topic is for tweets that discuss mask-wearing outside the context of COVID-19, such as masking to avoid breathing in volcanic ashfall, and for tweets with insufficient English content to accurately ascertain their stance towards universal masking. Each tweet can only be assigned one out of the four labels.

A number of factors influence the stance assigned to a tweet, and these are the explicit and implicit statements found within the tweet, non-body-text resources (images, videos, and hyperlinks), the conversation which the tweet is a part of (if any), the history of its author (timeline and profile page), and sincerity (humor and sarcasm). An explicit statement is one that states its position on universal public masking in a fairly clear manner, e.g., "please wear a mask". An implicit statement requires the coder to make a conjecture on its stance, e.g., expressing anger at people in shops not wearing masks typically implies that the tweet is supportive of universal public masking.

The primary goal of the labeling process is to obtain the most accurate stance for the tweets, so we leveraged resources that might not be accessible to a purely text-based machine learning classifier, e.g., images and videos which cannot be parsed, contents of hyperlinks' destinations which cannot be accessed. Another reason for using resources beyond a tweet and its immediate context is because those alone can be misleading or insufficient to determine the tweet's stance. An example of extra resources being helpful would be a tweet that appears supportive of mask-wearing, "don't forget to wear you mask when going out", that turns out to be anti-mask when one looks at the accompanying picture showing a person wearing a pair of panties on their face.

For tweets advocating for conditional masking, we evaluate their stances based on the tweets' intended impact on universal masking. Here are some examples. One commonly encountered type of tweet encourages anti-maskers to not wear masks and attend large gatherings, with the anti-maskers falling ill from COVID-19 often being the implied outcome. As these tweets' intended effect is likely to be scaring everyone into wearing masks by sincerely advocating for a segment of the population to not wear them, these tweets are classified as pro-mask. There are also tweets advocating for people to wear masks forever so that their ugly faces can be covered up. As these tweets intend to shame people into not wearing masks, they are classified as anti-mask. When a tweet advocates for public figures to wear/not wear masks or criticize public figures for wearing/not wearing masks, we examine the public figures' stances on wearing masks and the tweet author's own historical stance to determine the tweet's true stance. When criticism for not wearing a mask during some public event comes from a tweet author who has a history of anti-mask tweets and the criticism is directed towards a politician who has been an outspoken advocate for universal masking, we classify the criticism to be anti-mask.

A total of 3,010 tweets were manually assigned stance labels. This is the same set of randomly selected tweets whose theme labels were manually coded (see Section III-B3).

3) THEME IDENTIFICATION AND LABELING PROCEDURE

Unlike stance classification where a standard is first defined before labels are assigned to tweets, defining and labeling themes are concurrent processes. What remains the same as with the stance labeling process is the thorough use of all publicly available information to ensure the accuracy of the theme labels assigned to tweets.

We relied on the grounded theory/inductive coding process as described in [74]. In the first round/cycle of coding, concepts were identified through the textual data of the tweets as well as accompanying media (images, videos) from a sample of 3,000 randomly selected tweets from the dataset. As ten total duplicate instances originating from three unique tweets were found during the coding process, we sampled an additional ten random new and unique (non-intersecting) tweets. In the subsequent round of coding, concepts from the 3,010 tweets that demonstrated significant overlap with each other were merged into categories/themes. The final tally is 15 themes. Each tweet can have more than one theme with the exception of tweets that fall under either one of the two out-of-topic themes. Additionally, we picked one theme to be the ''main'' theme of a tweet. This main theme is *never* used in our analysis due to its low F1 scores but it is used to rectify the very rare issue of the machine learning classifier not assigning any theme at all to a tweet (Section III-B7).

The themes we have identified are listed below and their list numbering here corresponds to the numerical codes we assigned to them during coding:

1) **Declarative-personal**: Tweet author stating their stance on masking as it applies to themselves and/or people under their care (children), including policies they intend to enforce upon visiting guests. Tweets recounting personal observations, such as a Twitter user's description of the degree of compliance towards mask mandates at their neighborhood store, also falls under this theme.
2) **Authority-medical**: Tweets where medical authorities directly or indirectly exercise their authority. A medical authority may be an official Twitter account representing a healthcare agency or a scientific institution. The authority can also be an individual emphasizing their healthcare or scientific credentials when making claims about masking, including the deliberate use of technical jargon to appear more knowledgeable or credible. The veracity of the claims — whether they are scientific or pseudo-scientific — is immaterial. The focus of this theme is on tweets leveraging the public's trust in and deference to science (not as a process of knowledge generation but as a source of knowledge) and scientific authorities to compel others to wear or not wear masks.
3) **Encouragement**: Tweets actively encouraging others to wear masks or discouraging others from doing so. Encouragement must be explicit and not inferred. For instance, criticism of people not wearing masks alone is insufficient to count as encouragement.
4) **Appraisal-criticism**: Tweets criticizing or, very rarely, complimenting masking behaviors. This includes jokes, sarcastic remarks, and insults, e.g., mocking mask-wearers for being fearful of COVID-19 and lacking masculinity or mocking non-mask-wearing Trump for lacking masculinity and being afraid of masks ruining the orange cosmetic foundation he wears on his face.
5) **Concerns-effects**: Tweets discussing challenges, concerns, and side effects related to universal masking. The concerns can be about physiological (e.g., breathing difficulty) and psychological (e.g., PTSD) issues brought about by the act of wearing the mask itself. The concerns can also be about masking compliance and issues in procuring masks (cost, lack of supplies etc.). A tweet cannot be assigned this theme based solely on the presence of observations (e.g., tweet author noting that people do not wear masks in stores) within it. Tweet author must state their opinions on their observations (e.g., tweet author expressing fear of being infected by unmasked people in stores).
6) **Individualism-liberty-collectivism**: Tweets where the impact mask ordinances have over freedom, specifically personal freedom and individual liberty, is mentioned. Tweets assigned this theme most commonly emphasize either the need to put collective, communal, and group interests ahead of individual and personal interests or vice versa. Tweets discussing the freedom to wear masks (e.g., hospitals barring their workers from wearing masks) or not wear masks (e.g., non-masked individuals prohibited from entering stores) also fall under this theme. Another common manifestation of this theme are tweets talking about restrictions on freedom. This can be as simple as a tweet where the author celebrates not *having* to wear masks.
7) **Conspiracy**: Tweets mentioning conspiracies (actions hidden from the general public), such as masks harboring tracking chips or masks being secret signs of subservience to a new world order.
8) **Conditional**: Tweets discussing mask-wearing only in certain situations/circumstances or if the masks meet certain criteria. Examples are masking only when social distancing cannot be observed, masking only if the government allows for shops to reopen and matches in stadiums to be held, and masking only if the masks are of a better grade than normal surgical masks.
9) **Authority-non-medical**: Tweets leveraging non-medical authority to support their stances. The support requested can go beyond making an argument more persuasive, it can also be asking for the authorities to enact or enforce policies, such as asking store owners to ensure customers in stores wear masks. Some examples of the forms which this type of authority can take include political leaders, celebrities, organizations (businesses, countries), law enforcement (agencies, legal rights, legal documents such as a nation's constitution), and religion. The commonality shared by all these authorities is the absence of any ostensible scientific or healthcare credentials. The authority is

assumed to have a receptive audience outside of Twitter and/or some means to make others comply with it/its edicts.

10) **News**: Tweets sharing news articles, including tweets made from a media company's official account and journalists reporting news in the form of tweets.
11) **Mask-type-style**: Tweets advocating for certain types of masks. This could be people advocating for masks which are more effective, more comfortable, more stylish etc. It can also be commercial ads for masks and mask-related products.
12) **Questions-jokes-ambiguous-neutrality**: Tweets where only an ambivalent attitude towards masking can be detected. Personal observations and commentaries that do not lean either way in the masking debate count as having ambiguous neutrality. People asking similarly neutral questions, such as those concerning the safety of wearing masks long-term and the technicalities of being in full compliance with mask mandates, are also ambiguously neutral. Another subcategory are jokes that cannot be construed as being either in favor or against masking; jokes that can be will fall under the "appraisal-criticism" theme instead.
13) **Principled-neutrality**: Tweets that clearly state their neutrality on the mask-wearing debate, e.g., declaring that they do not know enough to advocate for or against masking.
14) **Out-of-topic (mask)**: Out-of-topic tweets discussing mask-wearing in non-COVID-19 contexts, such as air pollution from volcanic eruption, seasonal allergies, anonymity, movies, video games, and sexual purposes.
15) **Out-of-topic (language)**: A number of tweets mix English with a non-English language, most commonly Tagalog, Urdu, Hindi, or Malay, when discussing mask-wearing. Within the context of this research, Nigerian pidgin does not count as English. If the English portions of the tweets are insufficient for determining the stance and theme of the tweet, then they will be assigned this theme, even if they are clearly discussing masking as it relates to COVID-19.

4) DISTRIBUTION OF METRIC VALUES OVER TWEETS WITH MULTIPLE THEMES

Since our tweets can have more than one theme, the problem of splitting the share of the values of tweet metrics — likes, retweets, replies — between the themes naturally arises. For instance, if a tweet with *encourage* and *appraisal-criticize* themes received 20 likes, what percentage of those likes is attributable to each theme? There is no information that could guide us in giving each theme its correct share of the values. Equal distribution is our only option. As we felt that equally dividing the values excessively penalizes themes that are commonly found in the company of other themes instead of being alone, we opted to let all themes receive the full share of values. If a tweet received 20 likes, each theme associated with the theme is assumed to be responsible for all 20 likes. This attribution method inflates the metrics of themes but not tweets. In other words, tweets with multiple themes still have the same values for their metrics, but it does result in commonly occurring themes having larger values.

5) AUTOMATIC LABELING WITH DistilBERT

BERT (Bidirectional Encoder Representations from Transformers) belongs to the transformers class of neural network architecture which has proven adept at natural language processing tasks. Unsurprisingly, BERT has been used with great success in determining the stances of text sequences [59], [75], [76]. However, the formulation of stance classification tasks in these existing works are different from ours. In [76], the task is to determine if pairs of text sequences have the same stance towards an issue. In [59] and [75], the goal is to determine if a perspective (e.g., "global warming is a natural cyclical process") supports or opposes a given claim (e.g., "human-driven climate change is real"). Our work covers only one topic, which is fighting the COVID-19 pandemic though widespread mask-wearing, so there is effectively only one claim. Conditioning the transformer to determine stances based on a claim is unnecessary within our specific application.

The features used for classification are the username, name, date, mentions, and body text of a tweet. A username is unique while a name is a non-unique display name. Mentions are a list of usernames whom a tweet is being directed at. Names may contain text that hints at a person's stance (e.g., users in favor of masking have often appended "wear masks" to their display names) hence using them as a feature.

We used DistilBERT instead of BERT. DistilBERT runs 60% faster than BERT while still retaining 97% of its performance, which greatly speeds up the labeling of two million tweets. We modified DistilBERT so that each feature text sequence is enclosed with a `[CLS]` aggregator token in front and a `[SEP]` separator token at the end. Using the embedded representation from multiple CLS tokens gave better performance than using just one. We settled on this configuration after having tested other configurations, which were (1) concatenating all feature text sequences and enclosing them with only one pair of CLS and SEP and (2) using only a single CLS token at the start but appending a SEP token at the end of each feature text sequence.

Machine learning classifiers are essentially statistical pattern recognition algorithms and are not true artificial general intelligence (AGI). Therefore, they cannot actually understand the guidelines for labeling stances and themes that we outlined in earlier sections. The algorithms instead attempt to seek out patterns within the data that allow them to reliably reproduce the labeling results from human coders, regardless of whether humans relied on the pattern/information during the coding process.

6) LABELING PERFORMANCE

Theme-identification is a series of binary classification tasks as we allow for each tweet to have multiple themes assigned

TABLE 1. DistilBERT classification performance for stances, themes as a single label 15-class problem, and themes as multilabel binary class problem. Scores are means from the test splits of 10-fold cross validation after training for 5 epochs.

# classes	Label	Micro-F1	Macro-F1
4	Stance	0.5451	0.6980
15	Main theme	0.3078	0.5433
2	Theme 1 *declarative-personal*	0.8692	0.9123
2	Theme 2 *authority-medic*	0.7849	0.9003
2	Theme 3 *encourage*	0.8067	0.8266
2	Theme 4 *appraisal-criticize*	0.7917	0.7930
2	Theme 5 *concerns-effects*	0.6790	0.9030
2	Theme 6 *indiv-lib-collective*	0.7430	0.8319
2	Theme 7 *conspiracy*	0.4910	0.9648
2	Theme 8 *conditional*	0.5905	0.9037
2	Theme 9 *authority-non-medic*	0.7838	0.8691
2	Theme 10 *news*	0.8238	0.9671
2	Theme 11 *mask-type-style*	0.6358	0.9379
2	Theme 12 *question-joke-ambi*	0.5031	0.9322
2	Theme 13 *neutral-prinp*	0.4964	0.9857
2	Theme 14 *OOT-mask*	0.6297	0.9671
2	Theme 15 *OOT-language*	0.4987	0.9947

to it. Stance identification is a multiclass classification problem with four classes.

To determine the optimal number of training epochs for maximizing classification performance and to assess the generalizability of the model, we used a nested cross validation approach that used iterative stratification and relied on micro-F1 and macro-F1 scores for both loops. The scores of the validation sets of the 3-fold inner loop is used to determine the optimal number of epochs. Means of the scores for the test sets of the 10-fold outer loop reported in Table 1 estimate generalization error.

Note that prior to classifying the full 2M set of tweets, we retrained the classifier with all 3,010 hand-labeled tweets.

Different problems have a different optimal number of training epochs so we settled on 5 epochs as the best compromise. Our classifier's performance is comparable to the micro-F1 scores (average of 0.583 for [77]) and macro-F1 scores (average of 0.691 for [77], maximum of 0.7313 for [58] compared to our maximum of 0.7309) for tweet stance classification tasks reported in earlier works. Reference [58] is a single-claim three-class problem (pro-vaccine, anti-vaccine, and neutral) while [77] is a multiple-claim (multiple-target) three-class problem (support, against, and neither). Note that the computation of F1 in [77] is non-standard in that it "does not give any credit for correctly classifying 'neither' instances [but] the system has to correctly predict all three classes to avoid being penalized for misclassifying 'neither' instances as 'favor' or 'against'."

7) CLEANING LABELS

A very small number of machine predicted labels have issues. A tweet that has an out-of-topic stance cannot have a non-out-of-topic theme and vice versa. However, there are 5,407 tweets with OOT stances and non-OOT multilabel themes and there is also 1 tweet with a non-OOT stance and an OOT multilabel theme. Additionally, 616 tweets are assigned both OOT and non-OOT multilabel themes. 65,339 tweets were not given a single multilabel theme by the classifier. In total, these tweets with problematic labels constitute 3.40% of the dataset.

To remedy the issue of tweets not having any multilabel themes, we took the predicted main theme as the multilabel theme provided that doing so does not cause an OOT theme to be assigned to a tweet with a non-OOT stance or vice versa. This reduced the number of tweets without any multilabel themes down to 5,789. All 616 tweets with both OOT and non-OOT themes had their OOT-themes removed. We could not address the other issues as we are uncertain if the prediction error lies with the stance labels or the theme labels. As the remaining tweets with problematic labels total 11,197 or just 0.53% of the dataset, their exclusion should not affect our analyses significantly.

8) EXAMPLES

To give readers an insight into the stance and theme labeling process, we provide some examples here. The tweets found in these examples are in no way representative of the diversity of tweets that can be associated with these stances and themes.

"The airport is a private place of business that is glad to receive City taxpaying funds. The requirement to have to wear a mask is outrageous. Maybe @MayorGallego should stop giving millions to help them and see if they continue requiring people to wear them." This tweet has a negative stance and features the *appraisal-criticize* and *authority-non-medic* themes.

"When did I dispute mask wearing? I simply said people should be free to choose their own level of risk tolerance and decide where they want to go. If you have to go into a business and they require masks, you should wear it. But you should be allowed to take any risks you want". This tweet has a neutral stance and features the *indiv-lib-collective*, *authority-non-medic*, and *neutral-prinp* themes.

"Humans are strange. We all do basic things to survive, like eat and sleep, but those human who feel their rights are restricted or that they won't wear a mask because "I don't want to" - they are the reason our new normal is going to last so much longer. Thanks a lot, shitbags." This tweet has a positive stance and features the *appraisal-criticize* and *indiv-lib-collective* themes.

C. AUTOMATED SUMMARIES WITH BART

Taking inspiration from [48], where DistilBART[3] was used to summarize tweet clusters, we also used transformers to summarize groups of tweets for analysis. Key differences in our method are that the groups we are summarizing are manually defined and we did not use DistilBART trained on the XSum extreme summarization task. We instead used Facebook's original BART [78] trained on the CNN summarization task.

[3] BERT, BART, DistilBERT, and DistilBART are different but related types of transformer.

A desire for more accurate summariztion motivated this choice. As we did not need to generate a large number of summaries, the speed advantage offered by DistilBART over BART did not matter to us. We avoided using the weights from the XSum task because XSum summaries are highly abstractive, meaning that the summaries tend not to resemble the source sentences. In practice, this meant that there is a strong tendency for a transformer trained on XSum to "hallucinate" words and narratives that were absent in the seed tweets, giving us misleading summaries. In comparison, summaries in the CNN summarization task tend to resemble source sentences so a transformer trained on this task is much less inclined towards inventing new details.

To examine some phenomenon that interests us through BART summaries, we first filter tweets in our dataset so that only tweets representative of the phenomenon of interest remain. This filter can either be a specific time window, stances, themes, conversation IDs, or some combination of all of them. Usually the number of tweets that remain after filtering are too large to be accommodated by the transformer, which has a length limit on the seed text used to generate summaries. Therefore, a second round of filtering is almost always required to a select a smaller number of tweets with even greater representativeness, with the centroid proximity acting as the criterion this time. In the second round of filtering, BART is used to compute the sentence representation (encoded in the special <s> token) for all the tweets that remain after the first round of filtering. The top 30 tweets closest to the centroid of the sentence representations are selected to be the seed text for generating summaries.

Tweets selected as seed text are cleaned by removing hyperlinks, concatenated, and tokenized. If the number of tokens exceeded BART's 1024-input token limit, they are truncated to fit the limit. BART generates a summary with a maximum output length of 100 tokens from the input tokens.

The method of obtaining the top 30 tweets closest to a centroid was also used by us to study tweets representative of a conversation or of a particular day without generating a summary through BART.

D. LIKES AND RETWEETS AS PROXIES FOR THE SILENT MAJORITY

Studying the mask-wearing conversation purely through tweet text alone may give an inaccurate impression on the interests and leanings of Twitter users. This is because many people who register for social media platforms are often not prolific users or are users who engage with the platform in a low-visibility manner; these *lurkers* form a silent majority. Lurkers are not to be confused with churners, who register for a social media service but never use it. Research has indicated that, while not quite nine-tenths,[4] there are still three-quarters of Twitter users who are lurkers [79]. Not accounting for lurkers have been given as a reason for the failure in predicting election outcomes through polling Twitter sentiment [80].

Likes and retweets can offer a glimpse into where the silent majority stand on issues such as masking as both are low effort engagement activities, requiring but a tap or a click from the user. While not as clear-cut as explicit statements, liking a tweet is highly likely an unironic indicator of supporting the tweet's stance. Retweeting indicates that an idea expressed in a tweet is attention-worthy and/or cannot be ignored. If one agrees with a tweet, then positive attention is garnered through exposing the tweet to a sympathetic audience. If one disagrees with a tweet, then negative attention is gained through mockery or refutation by an unsympathetic audience. And if one is ambivalent, increasing visibility of a tweet can help foster debate.

1) RATIOMETRICS

Ardent users of Twitter have long noticed that disagreeable tweets tend to exhibit a certain tendency on the three metrics — replies, retweets, and likes — visible to the end user, which is that the more disliked by an audience a tweet is, the more the replies outnumber the likes or the retweets that the tweet receives [81], [82], [83]. This phenomenon is known colloquially among Twitter users as "being ratioed" and it has garnered sufficient attention that Merriam-Webster has put "ratio", "ratioed", and "ratioing" on the list of words that they are watching [82]. The idea of "ratio" being capable of revealing how a tweet is judged by its audience is gaining traction in scholarly circles as well. There has been a recent pioneering work that focuses solely on examining how ratios of Twitter metrics can be used to characterize tweets, using a case study of tweets from two highly polarizing political figures, Donald Trump and Barack Obama [84].

In our work we focus on three types of ratios (ratiometrics) — $N_{\text{retweets}} : N_{\text{likes}}$, $N_{\text{replies}} : N_{\text{likes}}$, and $N_{\text{replies}} : N_{\text{retweets}}$. The interpretations for the three ratios are discussed in Section IV-D5.a. We chose not to divide one raw number by another, e.g., $N_{\text{replies}}/N_{\text{likes}}$, to obtain the ratios because many tweets in our dataset have not received either a single like, retweet, or reply. An excessive number of tweets would have to be excluded due to the denominator being zero if we simply divided two metrics. To overcome this problem, all ratios are calculated using the following equation:

$$\begin{aligned} R_{\text{metric1-metric2}} &= N_{\text{metric1}} : N_{\text{metric2}} \\ &= \frac{N_{\text{metric1}} - N_{\text{metric2}}}{N_{\text{metric1}} + N_{\text{metric2}}} \end{aligned} \tag{1}$$

An additional advantage of this formulation is that it prevents extreme values from dominating the results compared to division by raw numbers or division with one added to the denominator (to avoid division by zero). We illustrate with an example featuring two tweets, tweet A with 1 reply and 100 likes and tweet B with 10 replies and 100 likes. The ratios are 1/100 and 1/10, and 1/10 for division by raw number, $1/101 \approx 0.009$ and $10/101 \approx 0.099$ for division with one added to the denominator, and -0.98 and -0.82 with Equation 1, respectively. The difference between the two tweets' ratios is relatively muted with Equation 1,

[4] Nielsen's rule.

but the other two ratio calculations greatly exaggerate (by an order of magnitude) the difference in ratios between the two tweets. Ratios are also better preserved with Equation 1 compared to adding one to the denominator. If C has 1 reply and 2 likes while D has 2 replies and 4 likes, their ratios should be equivalent, 1/2, but the method of adding one to the denominator changes them to 1/2 and 2/5. In contrast, Equation 1 keeps both ratios as −1/3.

Even with this formulation, there are still many tweets whose denominator terms sum to zero that need to be excluded from analysis. Only 1,247,174 tweets in total were used in our ratiometrics analysis, with retweets-likes, replies-likes, and replies-retweets ratios each having 1,058,102, 1,226,976, and 801,960 tweets respectively.

E. WORD IMPORTANCE

To discover which words were most important to different mask-wearing themes and stances, for each theme (or stance), we divided the sum over all tweets of each word's TF-IDF values by the number of times each word has appeared in the tweet corpus. The count for a word's appearance is boolean per tweet, e.g., a word appearing thrice in one tweet still counts as one. This translates to the following equation:

$$\frac{(\text{theme}_{a\times b})^{\intercal}\cdot \text{TF-IDF}_{a\times c}}{(\text{theme}_{a\times b})^{\intercal}\cdot A_{a\times c}}$$
$$\text{where } A = \begin{cases} 1 & \text{if term freq.} > 0 \\ 0 & \text{otherwise} \end{cases}$$
$$\text{for } a \text{ tweets, } b \text{ themes, and } c \text{ words} \qquad (2)$$

The equation above is the same for stances.

We selected only the top 60 words for analysis, with the ranking based on the sum of each word's TF-IDF values across all tweets. English stop words from NLTK and uninformative words, namely "wear", "wearing", "mask", "masks", "https", "www", "twitter", "com", "pic", "status", were excluded from our analysis. The uninformative words are either different ways of expressing the key words we used to build our dateset ("mask" and "wear") or they are words that make up the hyperlinks commonly found in tweets, including links to pictures.

IV. ANALYSIS

A. CO-OCCURRENCE OF THEMES

Figure 1 shows how frequently different themes are found within the same tweet.

When a tweet is appraising the masking behavior of others (*appraisal-criticize*), two other themes commonly appear alongside it, which are those that raise the issue of individual freedoms or collective interests (*indiv-lib-collective*) and those that encourage others to wear or not wear masks (*encourage*). A synthetic example of a tweet embodying all three themes, based on our experience coding tweets: "People refusing to mask up want to kill grandmas. Wear your mask!"

Tweets invoking medical authorities (*authority-medic*) are strongly associated with almost all themes except for those asking questions, joking, or expressing neutrality (*question-joke-ambi* and *neutral-prinp*). Tweets concerned with non-medical authorities that have the ability to influence universal masking compliance such as celebrities, politicians, law enforcement agencies, and department stores (*authority-non-medic*) also have a strong association with many other themes. Tweets with a *conspiracy* theme is a notable exception, as they are a little less likely to bring up *authority-non-medic* than *authority-medic*.

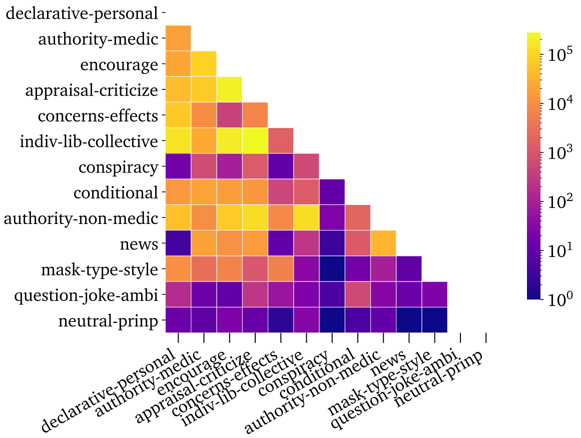


FIGURE 1. Counts of co-occurring themes. Color scale is on a log axis.

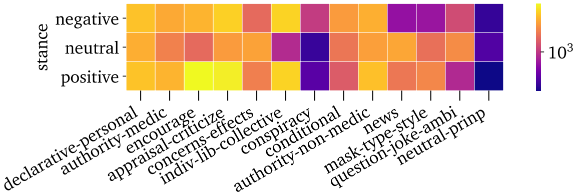


FIGURE 2. The number of tweets belonging to all possible combinations of stances and themes, excluding the out-of-topic stance and themes. Color scale is on a log axis.

News coverage being shared in tweets are rarely those that relate to the side effects of universal masking (*concerns-effects*) such as breathing difficulties. News regarding situations where masks can selectively be not worn (*conditional*) are more likely to be found than *concerns-effects*.

B. CO-OCCURRENCE OF THEMES AND STANCES

Figure 2 shows the counts of tweets for different combinations of stances and themes. The most commonly occurring types of tweet in our dataset are those encouraging people to wear masks (positive and *encourage*) and those criticizing, praising, or giving appraisal to the mask-wearing habits of other people in a way that advances an overall pro-mask agenda (positive and *appraisal-criticize*).

Certain themes occur far less frequently among tweets bearing a particular stance towards universal masking. The most prominent example is the *indiv-lib-collective* theme that is commonly found among tweets that raise the issue of personal freedom in choosing to either wear or not wear masks.

Few neutral tweets are associated with *indiv-lib-collective*; the bulk of the tweets with the *indiv-lib-collective* theme are evenly split between negative and neutral stances. Discussion or promotion of mask types and styles (*mask-type-style*) are also much more common among tweets which either support or are neutral towards universal masking. Presumably, those who are against masking would not bother themselves with the issue of picking the right mask to wear. Tweets featuring *news* are much less frequently associated with a negative stance towards masking, perhaps reflecting a distrust or disagreement towards universal masking position reported or promoted by mass media sources.

Differences in the distribution of stances are less stark with other themes. The number of tweets discussing mask-wearing on a conditional basis (*conditional*) becomes increasingly common the less positive a tweet is about universal masking. Tweets expressing concerns over issues related to universal masking (*concerns-effects*) are less likely to express either support or opposition to universal masking at the same time.

The findings in this section apply to the distribution of the number of likes, retweets, and replies received by tweets of a particular theme and stance combination as well because of the high Pearson's correlation coefficients between the counts of tweets, likes, retweets, and replies, which range from 0.9626 to 0.9958 (p-values range from 4.7×10^{-40} to 1.4×10^{-22}) for all 6 unique combinations of the different counts. As an example, the two most commonly posted types of tweets, positive & *encourage* and positive & *appraisal-criticize*, are also the types receiving the most likes, retweets, and replies.

C. WORD IMPORTANCE

Figure 3 shows which words are important to each theme of mask-wearing tweets based on aggregated TF-IDF scores.

The expletive ''fucking'' is important to tweets of all themes but especially with the *encourage* theme, which accords with our experience during the coding process where variations of ''wear a fucking mask'' are commonly encountered. The word ''sick'' is important to tweets with the *conditional* and *authority-medic* theme, reflecting the many tweets using scientific/medical authorities to bolster the claim that people should only wear masks if they are sick or caring for someone who is sick and not wear them if they are healthy. A slightly less important word for the *conditional* theme is ''outside'', which is often used by people proclaiming that masks are for indoors use only and should not be worn in open-air areas. The *question-joke-ambi* theme have multiple words that are important to it, likely due to the tweets bearing this theme being highly diverse. From examining a random sampling of tweets featuring the theme, the importance of the words ''public'' and ''work'' is due to neutral questions or pronouncements about having to wear masks in public and/or at work. The word ''glove'' is likely due to people uncertain about the necessity of wearing gloves alongside masks.

When comparing word importance scores purely across stances, we observe that the expletive ''fucking'' is important for all stances but most of all the positive stance. The word ''sick'' is more important to tweets with a negative stance than other stances and based on our examination of the corpus, the rhetoric of such tweets focus on masks not being able to prevent the getting sick or spreading sickness, reserving mask-wearing only for the sick, and the sick staying home negating the need for masks. The word ''please'' and ''always'' is important to positive tweets relative to other stances due to tweets encouraging others to mask up forming a large part of the positive-stance corpus. For neutral tweets, the relatively important words are ''really'', ''work'', ''day'', and ''gloves'', which are similar to those for *question-joke-ambi*, reflecting neutral tweets' tendency to feature people describing experiences wearing masks and sometimes gloves all day on their job and/or questioning the necessity of doing so.

A number of other keywords have low importance scores across all themes but still scored high enough to place among the top 60, which shows that they are important overall to the tweet collection even though they are not important to any particular theme within the context of the top 60 words, and these words are: ''coronavirus'', ''virus'', ''social'', ''distancing'', ''please'', ''stay'', ''home'', ''keep'', ''safe'', ''take'', ''care'', ''protect'', ''health'', ''wash'', and ''hands''. Their presence validate our data collection strategy, as they turned up even though we did not use any COVID-19 or healthcare related keywords. An additional word identified as important to many tweets is ''trump'', demonstrating both the US-centric nature of our dataset and the centrality of President Trump to the mask-wearing discourse.

D. TEMPORAL TRENDS

1) JUNE ANOMALY

Plotting the counts of tweets, N_{tweets}, over time (Figure 6) revealed that a massive drop in volume occurred in a short period of time spanning the end of of May (approximately May 26, 2020) and the start of June (approximately June 8, 2020). For convenience's sake, we shall refer to this event and this time period as the *June anomaly*. We call attention to the June anomaly here because we will be constantly referencing to it due to it altering many aspects of the prevailing conversation dynamic.

The possibility that the June anomaly is a result of some data collection error can be ruled out due to the existence of another, earlier paper studying COVID-19 mask-wearing conversation [48] which showed a similar drop in their plot of tweet count. Crucially, this drop in tweet volume was recorded in spite of a different data collection strategy (coronavirus-related tweets were first collected before further filtering through mask-related keyphrases), a different API (Streaming API), a different set of keyphrases, and covering a different time period (March 17, 2020 to July 27, 2020).

2) PROPORTIONS OF STANCES AND THEMES

Tweets with a negative stance and the *appraisal-criticize* theme exhibit a diurnal pattern, as can be seen in the midday

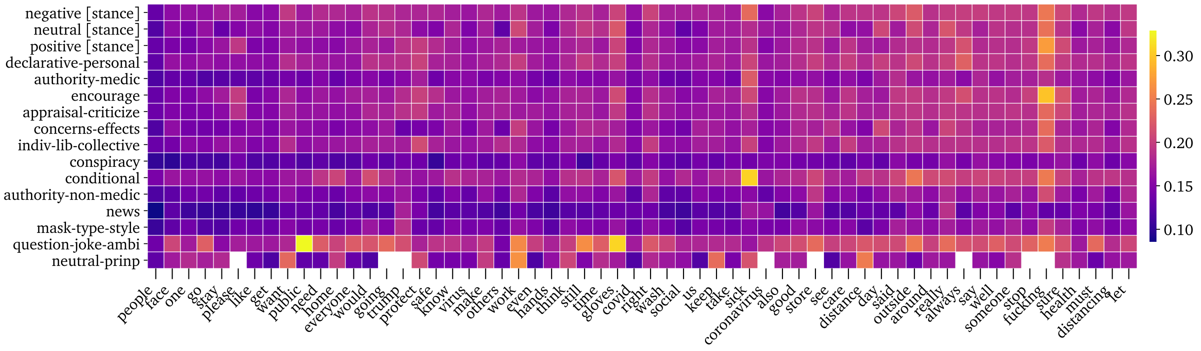


FIGURE 3. Word importance to different stances and themes of mask-wearing tweets.

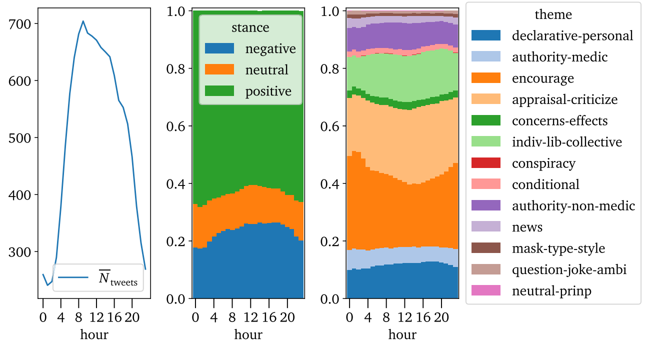


FIGURE 4. Left: mean tweet count per hour. Middle: per hour proportions of stances towards universal masking averaged (mean) over all days. Right: per hour proportions of mask-related tweets' themes averaged (mean) over all days. Timezone used is Pacific Daylight Time.

bulges on both stacked bar plots in Figure 4. Late in the night and early in the morning, the proportions of tweets with the aforementioned stance and theme shrink, with the lost shares taken up by tweets with positive stance and *encourage* theme. The proportions for the remaining stance and themes remain mostly unchanged. Based on these patterns and the assumption that most tweets during the pandemic are posted during each region's non-nocturnal hours,[5] the opposition towards masking as well as the harshest criticism for masking behavior appears to be stronger in the Western Hemisphere — most likely the US based on the language of our dataset as well as the celebrities and politicians regularly mentioned in tweets — than other parts of the world.

Figure 5 portrays the changes in proportions of stances and themes at a much longer time scale of months. Unlike Figure 4, Figure 5 includes the out-of-topic stance and themes to demonstrate that (1) tweets using a mix of English and non-English language (*OOT-language*) were never a significant minority over the entire dataset's timespan and (2) the share of tweets that discuss masking in a context other than COVID-19 (*OOT-mask*) and out-of-topic tweets in general (*stance-OOT*, *OOT-language*, *OOT-mask*) virtually disappeared after March.

The proportions plotted in Figure 5 allow us to see sudden growth in tweet volume that might be less obvious on raw count plots plotted in Figure 6. Take for instance the period between January 12 and January 14, which showed ephemeral spikes in positive stance and *encourage* theme in Figure 5. We shall discuss the January 12–14 period in greater detail because it reveals a weakness in our data filtering and classification strategy. Examining the four conversations with the highest tweet counts in that period, we found that they were made up of three standout events, none of which were related to COVID-19. Two of the events could easily be mistaken to be COVID-19 related: fans asking Trump supporter Scott Presler to wear a mask to shield himself from being infected by airborne diseases carried by homeless people and actress Jameela Jamil expressing concern over catching the seasonal flu from recycled air on planes. The third event is the January 12 volcano eruption in the Philippines, which led to many tweets calling for masks to be worn that did not contain words to indicate their relationship to the eruption, e.g, "ashfall". Out of the 73 tweets that make up the four conversations, only 20 were assigned an out-of-topic label.

There was another peak of tweets with positive stance and *encourage* theme from the end of January to early February. This time, based on our examination of news articles and the six largest Twitter conversations from this time period, there is much less uncertainty on whether there are misclassifications of out-of-topic tweets; they are not misclassifications. One conversation consists of fans of Filipino boy band SB19 wishing JoshTin well and asking him to wear a mask after that person has fallen ill. Another three conversations were asking for Korean pop stars Cho Seungyon, Rocky, and Jaehyun to wear masks. The fifth conversation along the same vein featured a Chinese pop star named Jackson Wang. Although it is possible that these requests could be in response to seasonal flu, Korean news articles[6] from this time period spoke of a new trend in the entertainment industry encouraging mask-wearing in response to the coronavirus, ruling

[5] Plots of Twitter activity for cities around the world, including two American ones [85], [86], show that activity peaks twice in a day, once in the morning and once more in the evening. Activity is the lowest between midnight and dawn.

[6] e.g., https://n.news.naver.com/article/088/0000630884

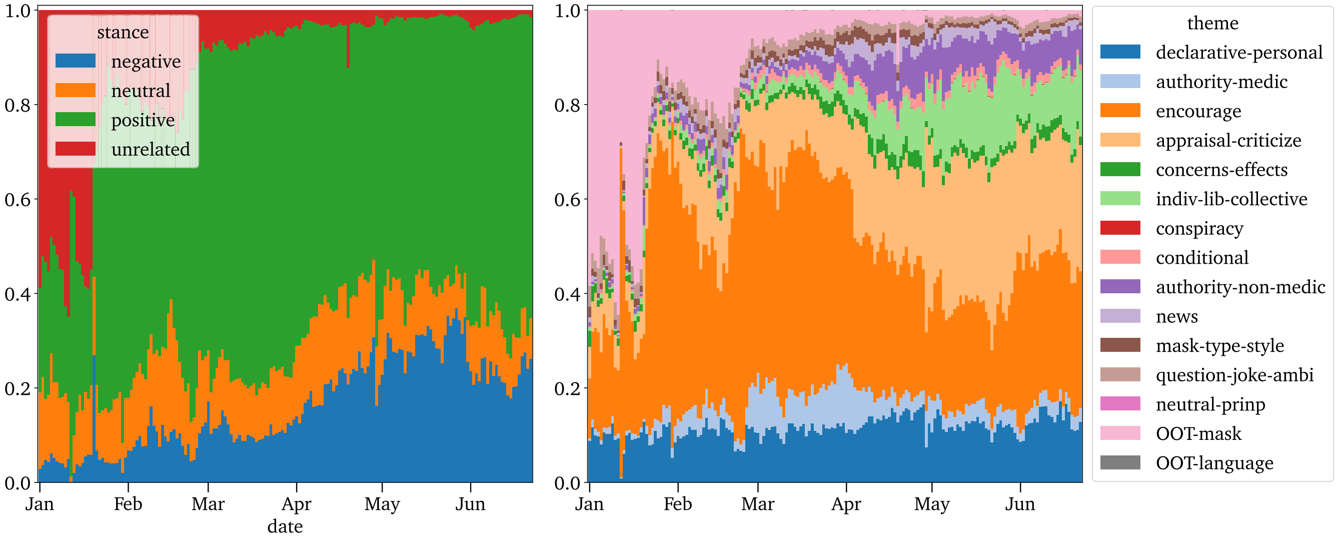


FIGURE 5. Left: temporal evolution of the proportions of stances towards universal masking. Right: temporal evolution of the proportions of mask-related tweets' themes.

out the seasonal flu possibility. Efforts to evacuate citizens of other countries from Wuhan, China also began in earnest near the end of January[7] and the pandemic was sufficiently threatening that news outlets have begun publishing special features on coronavirus,[8] further lessening the probability that the calls to wear masks was due to seasonal flu instead of COVID-19. The sixth and final conversation, which consisted of people defending their decision to wear a mask due to being perceived as sick and thus ostracized, was unambiguously related to the coronavirus because it featured keywords such as "coronavirus", "corona virus", "coronovirus", and "the virus".

Our decision to concentrate on analyzing tweets after March 1, 2020 is partially motivated by the potential for misclassification of tweets in earlier time periods as being COVID-19-related; the issue is not a concern post-March 1 based on our analysis of a random sampling of conversations after that date.

Early March saw a peak in proportions of positive stance and *encourage*. The conversations responsible for the peak will be elaborated upon in Section IV-D3 so we will withhold from discussing them here.

We now turn to the overall prevailing trends in proportions of stances and themes from early March all the way until the advent of the June anomaly. In this time period, the proportion of tweets ambivalent towards universal masking have stayed mostly the same. The share of tweets supportive of masking, while remaining a majority, withered slightly over time. Tweets against masking though has seen near-uninterrupted growth. In terms of themes, this period saw the rapid growth of *appraisal-criticize*, *indiv-lib-collective*, and *authority-non-medic* themes. Their growth came at the expense of the share of *encourage* and *authority-medic* themes, whose mostly stable proportions in March entered a constant state of decline in April and May. The change in proportions of the themes can perhaps be explained as people growing increasingly tired of admonitions to wear masks, a disinterest or perhaps distrust in advice from medical authorities, a growing interest in criticizing others for their stance on masking, and a growing concern over the enforcement of masking policies and what was believed to be the violation of personal freedom with the introduction of public masking requirements.

The June anomaly saw a decline in the shares of tweets with negative and neutral stances towards masking, with those shares taken up by tweets with positive stance. The proportions of themes found among tweets were also altered. Themes that were previously growing in size, *appraisal-criticize* and *indiv-lib-collective*, now entered into a state of decline. Meanwhile, *encourage* tweets grew its share, reversing its previous trend of decline.

While Twitter may not be the most accurate proxy for the American public due to its global nature, potential censorship (Section V-B), and self-selecting membership/participation, the proportions of stances actually concur quite well with public polls on mask-wearing stances. For instance, a survey conducted by the survey and market research firm SSRS for the Commonwealth Fund from May 13 to June 2, 2020, and involving 2,271 respondents found that "85% of adults believe that it is very or somewhat important to require everyone to wear a face mask 'at work, when shopping, and on public transportation'."[9]

3) COUNTS, STANCE MEANS, STANCE STANDARD DEVIATIONS

The left half of Figure 6 shows the temporal evolution of the counts and ratios of tweets, conversations, and users. A con-

[7] https://www.cnn.com/2020/01/29/world/wuhan-coronavirus-evacuations-intl/index.html

[8] e.g., https://www.cnn.com/asia/live-news/coronavirus-outbreak-01-29-20-intl-hnk/index.html

[9] https://www.medscape.com/viewarticle/933313

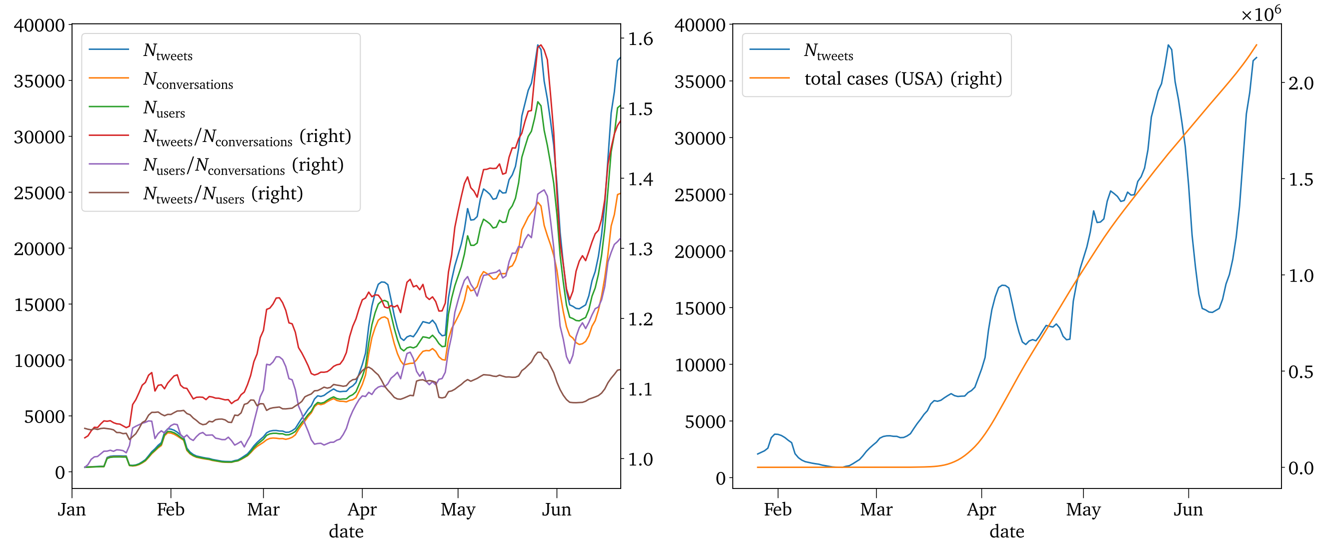


FIGURE 6. Left: seven-day rolling average of per day counts of tweets, counts of conversations (tweet threads), counts of tweets divided by counts of conversations, counts of users divided by counts of conversations, and counts of tweets divided by counts of conversations. Right: seven-day rolling average of per day counts of tweets with the cumulative total number of coronavirus cases in the US overlaid for comparison.

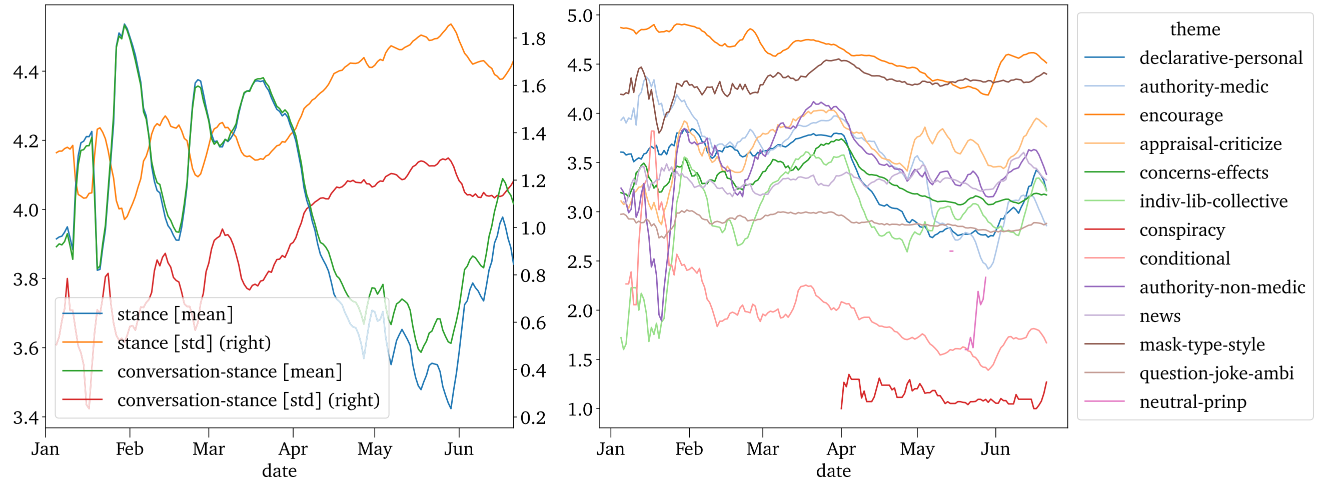


FIGURE 7. Left: seven-day rolling average of per day averages of tweet stances is represented by stance [mean]. Out-of-topic tweets were excluded. Numerical values of 1, 3, 5 stand for negative, neutral, and positive stances. Tweet stances averaged on a per conversation basis, then on a per day basis, is represented by conversation-stance [mean]. Standard deviations of the averages are included in the plot as well. Right: seven-day rolling average of per day averages of tweet stances grouped by themes.

versation, also known as a reply thread, consists of a tweet and its replies. $N_{\text{users}}/N_{\text{conversations}}$ represents the ratio of the number of users discussing masking to the number of unique conversations on masking. $N_{\text{tweets}}/N_{\text{conversations}}$ represents for the ratio of masking tweets to conversations. $N_{\text{tweets}}/N_{\text{users}}$ represents the ratio of tweets to users.

Based on the N_{tweets} line, we see that discussing mask-wearing was not much of a popular topic on Twitter in the early weeks of January, averaging less than 1,000 tweets on most days. In response to growing awareness of a coronavirus outbreak, as discussed earlier in Section IV-D2, there was a rash of tweets urging celebrities to wear mask in late January and early February before interest level dropped off again. Late February was when mask-wearing tweet count began growing almost daily, and late February coincided with the time period when COVID-19 cases were first detected in North America and Europe. At the beginning of May, there was a spike in the count of daily tweets discussing mask-wearing. The middle of May saw another sharp increase in tweet count. These two volume spikes happened despite declining daily (not cumulative) case count for the two major Anglophone countries, the US and the UK, throughout the entire month of May. The daily tweet count reached a peak of more than 350,000 at the end of May then abruptly dropped off over the next few days to around 150,000 — this is the event which we have been referring to as the June anomaly. So severe is this decline that all the growth in tweet count for the entire month of May was undone, returning to the level found at the end April. The conclusion of the June anomaly is followed by a quick recovery where the number of tweets returned to the same figure as the peak achieved at the very end of May.

Throughout the entire time range of our dataset, the shape of the $N_{\text{users}}/N_{\text{conversations}}$ line often closely follows that of $N_{\text{tweets}}/N_{\text{conversations}}$. However, $N_{\text{tweets}}/N_{\text{users}}$ did not change as much and therefore has a much flatter line. This indicates that the entry or exit of unique users every day is the primary driver behind the changes in the number of tweets found within a single conversation ($N_{\text{users}}/N_{\text{conversations}}$), not the same set of users increasing the intensity of their conversation ($N_{\text{tweets}}/N_{\text{users}}$).

When there are peaks in $N_{\text{users}}/N_{\text{conversations}}$ and $N_{\text{tweets}}/N_{\text{conversations}}$ but N_{users}, $N_{\text{conversations}}$, and N_{tweets} stay level, it indicates the presence of "lightning rod" conversations that attracts and accumulates attention. Early March serves as an illustrative example of such a situation. Examining the six conversations with the highest tweet counts around early March revealed that the gradual rise and drop off in interest on masking in this period can be attributed to a confluence of three events: Korean pop fans asking their idols to take better care of themselves by wearing masks (Korean pop fans' presence on English Twitter is significant enough to engage and succeed in vigilante acts such as taking over racist hashtags [87]), Republican Matt Gaetz's decision to wear a gas mask on the House floor drawing criticism, and debates over the ability of masks to prevent the coronavirus infection. If we look at the left half of Figure 7, we can see that as more participants joined these conversations, it had an overall effect of lowering the mean (less support for masking) and raising the standard deviation (making masking more contentious) of stances over all tweets and and also when taking the average of the averages in conversations.

Around early April, a situation that is the inverse of early March's occurred. $N_{\text{tweets}}/N_{\text{conversations}}$ and $N_{\text{users}}/N_{\text{conversations}}$ declined slightly but N_{users}, $N_{\text{conversations}}$, and N_{tweets} peaked. Users are dispersed into multiple conversations instead of being concentrated into a few large ones. Examining the raw data instead of the running averages showed that the peaks in tweet volume are on April 3 and 4. The largest conversation on April 3 was criticizing someone because their husband refuses to wear a mask. Looking at all tweets in April 3 and 4 without delving into any specific conversations showed a lot of generic tweets calling for people to wear mask, but we also noticed criticisms directed at a certain nebulous "he", whom we suspect to be President Trump. Our suspicion was confirmed when examining The Guardian's coronavirus news summary for April 4, which mentioned Trump ignoring US health officials' advice to people to cover their faces when outside. The dispersed influx of users and tweets however did not change the stance means and stance standard deviations, which had been trending downwards and upwards respectively since the end of March.

To summarize what we have learned thus far: concentrated attention without increased numbers of users and tweets can cause a shift in mean stance while dispersed attention with increased numbers of users and tweets appears incapable of changing mean stance.

We now explore an additional scenario: an abrupt and simultaneous dispersal of attention and reduction in user and tweet counts. We speak of the June anomaly, which stretches from the end of May to early June. The impact on stance mean is immediately observable; its slow decline was reversed immediately and returned to a local maximum that is comparable to the value at mid-April (left half of Figure 7). A reduction in the diversity of opinions is evinced by a drop in the stance standard deviation, which had been trending upwards since mid-March. The June anomaly also caused a number of themes whose mean stances were declining to reverse their trends, most notably *encourage*, *authority-non-medic*, and *authority-medic* (right half of Figure 7). The inflection points for these themes were before WHO changed their guidelines in June 5 so it cannot be attributed to the changed guideline.

Although we have previously relied upon the method of examining conversations and tweets from a particular time period to find the potential causes for increases in tweet counts, the same method cannot be used to satisfactorily explain decreases. The presence of tweets is insufficient for explaining the absence of user participation. So, at this point, we can only offer some speculations. One possibility is that COVID-19 is no longer deemed to be a threat and/or people have made peace with their position on universal masking. As a result, people no longer felt the need to proselytize, engage in vigorous apologia for their own position, and/or attack others who did not adhere to a similar position. This was ruled out because after the June anomaly ended, N_{users}, $N_{\text{conversations}}$, and N_{tweets} rebounded swiftly to pre-anomaly levels. We did not collect data beyond June 21, but the plots in [48] showed that counts for mask-related tweets increased up until the end of their data collection period, which was July 27.

The stance means for the *indiv-lib-collective*, *authority-non-medic*, and *appraisal-criticize* themes exhibit highly similar patterns (right half of Figure 7). Their mean stances all peaked in late March. Examining the tweets bearing all three themes in that time period showed that this was a result of a rise in complaints about store employees not being given the freedom to wear masks to protect themselves in stores. A month later in late April, their mean stances have all dropped. When we examine the tweets in late April, we saw that another narrative featuring all three themes, one that perhaps has greater appeal than store employee not being able to wear masks, has taken hold. Interestingly, this new narrative still revolves around stores, but this time around, it is the customers who are criticizing stores for infringing upon their personal rights by requiring them to wear masks.

4) COVID-19 STATISTICS

Case counts, deaths, and other COVID-19 statistics are often reported in news media and used by proponents of masking to encourage masking. The objective of this section is to

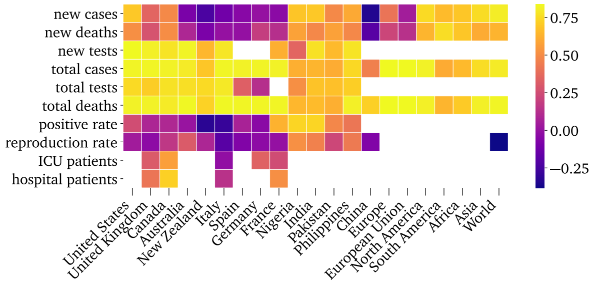


FIGURE 8. Pearson's correlation coefficients between the daily count of mask-related tweets and different daily coronavirus pandemic statistics for various countries and regions.

ascertain if the mask-wearing discussion on Twitter mirrors the "local" development of the pandemic, if they are driven more by the global pandemic situation, or neither. "Local" in this section refers to the US. The prevalence of US-centric entities in the tweet conversations we have examined and the cumulative per hour tweet count seen in Figure 4 strongly suggests that Americans, or at least people residing in the Western Hemisphere, are responsible for the vast majority of tweets in our dataset. To compare "local" with non-local, we also looked at the statistics for core Anglosphere countries, countries with large populations of English speakers (whose native languages are coincidentally the ones most commonly encountered for the *OOT-language* theme), non-Anglophone countries, and regions of the world. Our COVID-19 statistics came from Our World in Data (OWID)[10],[11] and OWID in turn obtained their data from various sources such as official reports released by countries and the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).

Figure 8 shows the Pearson's correlation coefficients between the daily tweet count and COVID-19 statistics for various countries and regions. For all countries and regions, total cases and total deaths are the statistics that consistently have high correlations with the tweet count. More obscure statistics, i.e. those that are not regularly reported in news media or reported by countries as evinced by the blanks space in Figure 8, tend not to correlate well with the tweet count. While many of the US's COVID-19 statistics for the US have good correlations with the tweet count, they are not exceptionally good compared to other Anglophone countries. Italy, a non-Anglophone country, have stronger correlations than the US for total cases, total tests, and total deaths. New case counts for Nigeria, Asia, and the world have stronger correlations with the daily tweet count than the US's daily new case count.

The June anomaly is not responsible for the low correlations. If we move back the time window to May 21, 2020 when calculating correlation coefficients, the correlations do increase for many countries and regions, but the patterns we have discussed earlier still hold.

The right half of Figure 6 depicts the counts of tweets with different stances alongside the US's total COVID-19 case counts. Tweet count and case count lines both follow a general upward trend but they are not close matches.

All in all, the evidence is weak for the tweet count to be directly influenced by the local, i.e. the US, pandemic situation as represented by statistics such as case and death counts. Furthermore, in our examinations of various conversations in Sections IV-D2 and IV-D3, case counts were never explicitly mentioned in them. Therefore, we believe that the likelier driver for new tweets discussing mask-wearing are the actions taken in response to the threat of the pandemic and the chain of reactions towards those actions (e.g., mask mandates, debate about mask mandates, people flouting mask mandates, debates about rule breakers).

5) RATIOMETRICS AND COUNTS OF LIKES, RETWEETS, AND REPLIES

Figures 9 and 11 shows change over time of tweet metrics' ratios, with the former figure broken down by stances and the latter by themes. Figure 10 shows change over time of tweet metrics for different stances while figure 12 does the same for different themes. Figure 13 shows the daily totals and averages of tweet metrics. As daily means of the counts for likes, retweets, and replies are prone to being influenced by outliers, even after smoothing with seven-day rolling averages, their plots tend to be bursty, making it difficult to extract meaningful patterns from them. Our discussion in this section will therefore focus more on the ratio plots, which feature normalized values.

a: INTERPRETING RATIOS

In short: controversial to mildly offensive tweets have positive $R_{\text{retweets-likes}}$, mildly offensive to very disagreeable tweets have positive $R_{\text{replies-retweets}}$, and very disagreeable tweets have positive $R_{\text{replies-retweets}}$.

We will work out how we arrived at these interpretations below.

In general, most tweets exhibit the following pattern for their metrics: $N_{\text{likes}} > N_{\text{retweets}} > N_{\text{replies}}$. In our dataset, based on the means, the ratio is approximately $55 : 13 : 4$. Ratiometrics are meant to highlight tweets deviating from this pattern. $N_{\text{replies}} : N_{\text{likes}}$ is the most commonly used ratiometric. If $N_{\text{replies}} > N_{\text{likes}}$ for a tweet, then that tweet is likely to be considered by its audience to be disagreeable or bad. This is a widely accepted definition of being "ratioed" on Twitter (Section III-D1). Only 221,214 tweets (10.53%) in our dataset meet the criteria of being ratioed, with the difference in magnitude between N_{replies} and N_{likes} within that small selection being just 1 at the 75th percentile. Using our ratio equation, $\frac{(N_{\text{replies}}-N_{\text{likes}})}{(N_{\text{replies}}+N_{\text{likes}})}$, increasingly positive values for the ratio $R_{\text{replies-likes}}$ corresponds to a tweet being increasingly disagreeable.

We offer two possible explanations for why replies outnumbering likes makes a tweet disagreeable. First, replying to

[10] https://github.com/owid/covid-19-data/tree/master/public/data

[11] https://covid.ourworldindata.org/data/owid-covid-data.csv

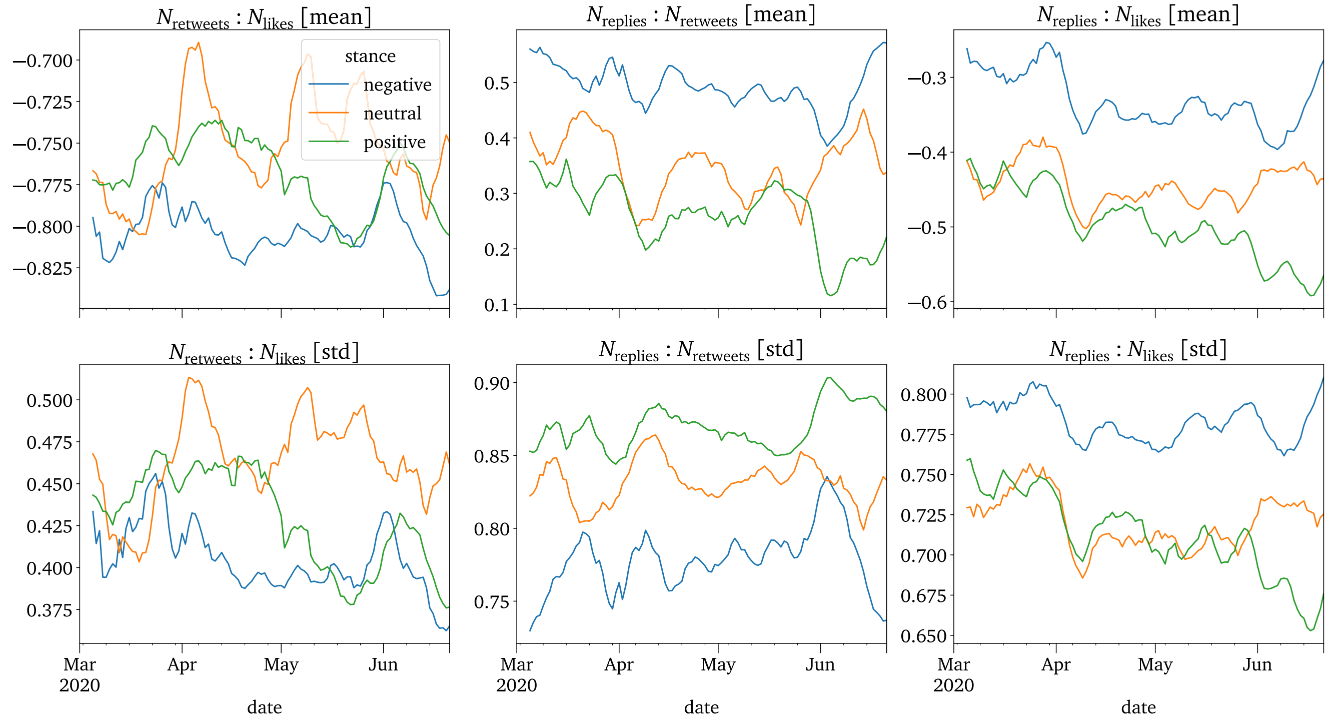


FIGURE 9. Upper row: seven-day rolling averages of various ratios between likes, retweets, and replies received by tweets, grouped by the tweets' stances. Lower row: standard deviations of the averages.

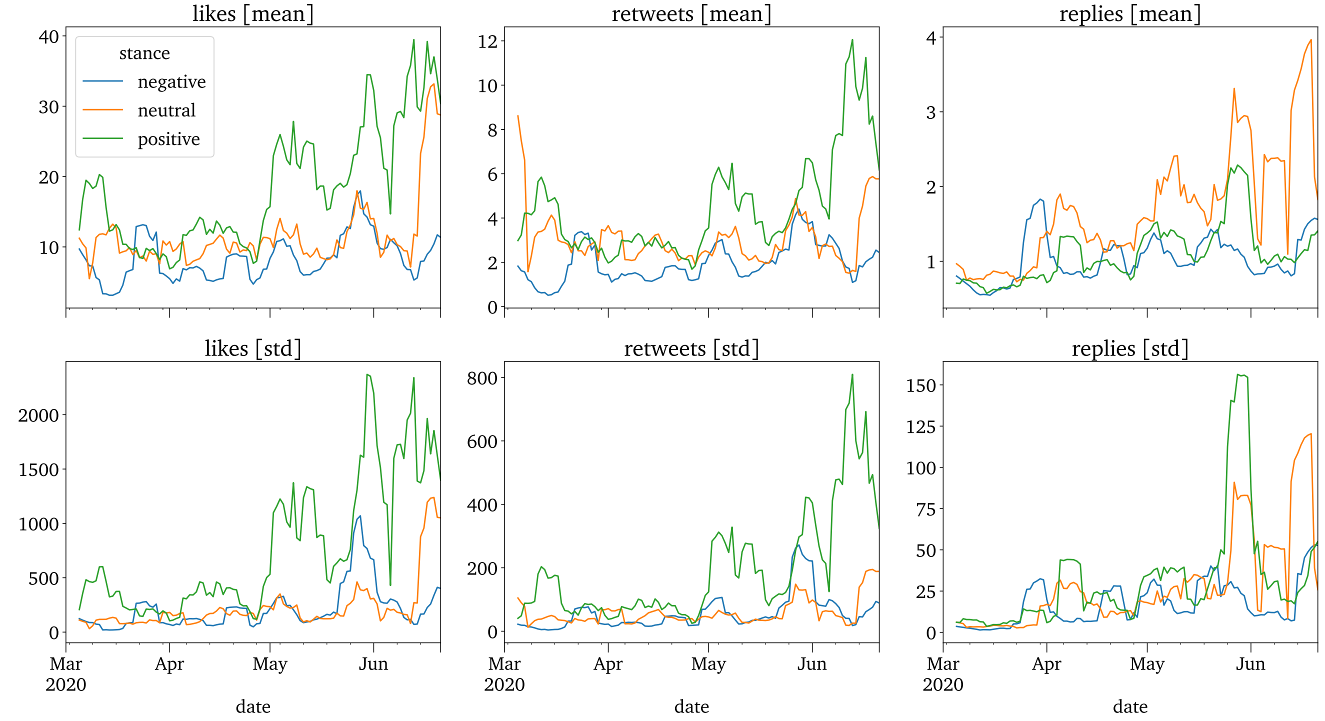


FIGURE 10. Upper row: seven-day rolling averages of the counts for likes, retweets, and replies received by tweets, grouped by the tweets' stances. Lower row: standard deviations of the averages.

a tweet is the only way to clearly express disagreement with a tweet because the only other interaction options available to a user are to like or to retweet the tweet. Second, typing out replies is a high-effort method of engaging with a tweet compared to simply ignoring a tweet, making replying a stronger expression of disapproval than ignoring a tweet. Not

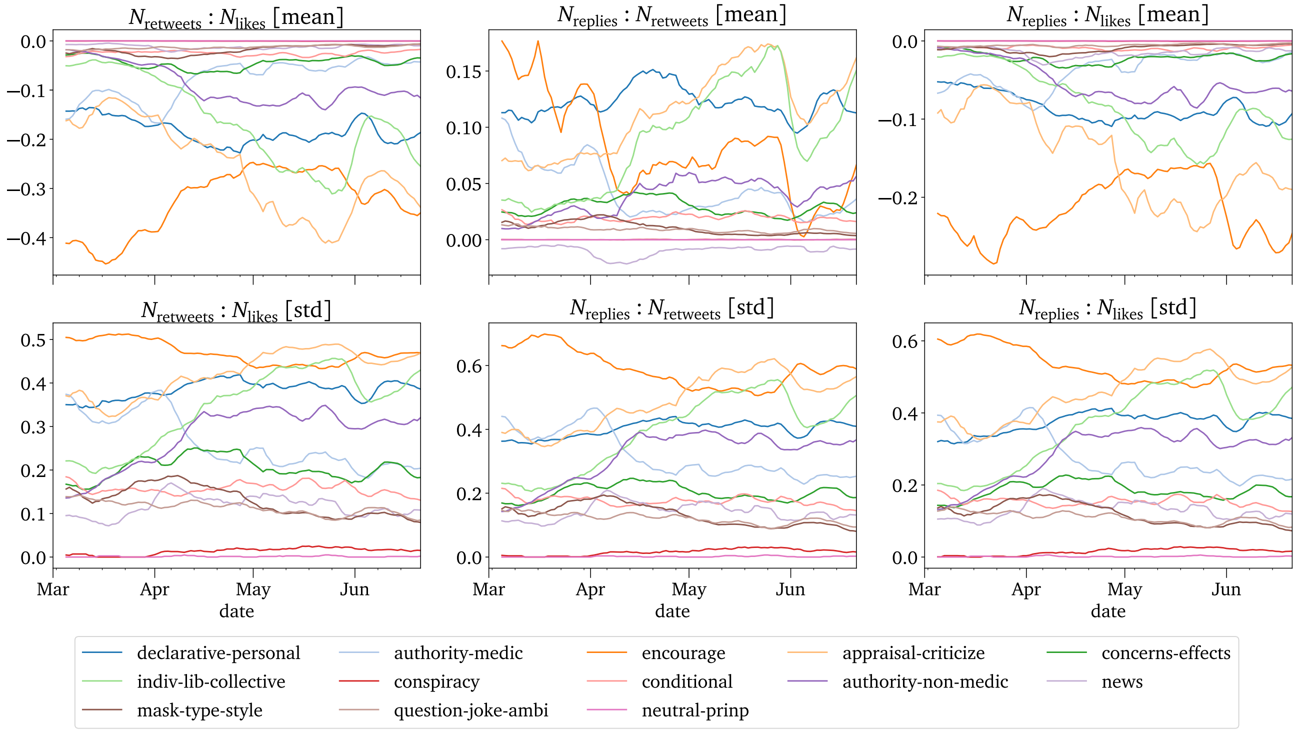


FIGURE 11. Upper row: seven-day rolling averages of various ratios between likes, retweets, and replies received by tweets with different themes. Lower row: standard deviations of the averages.

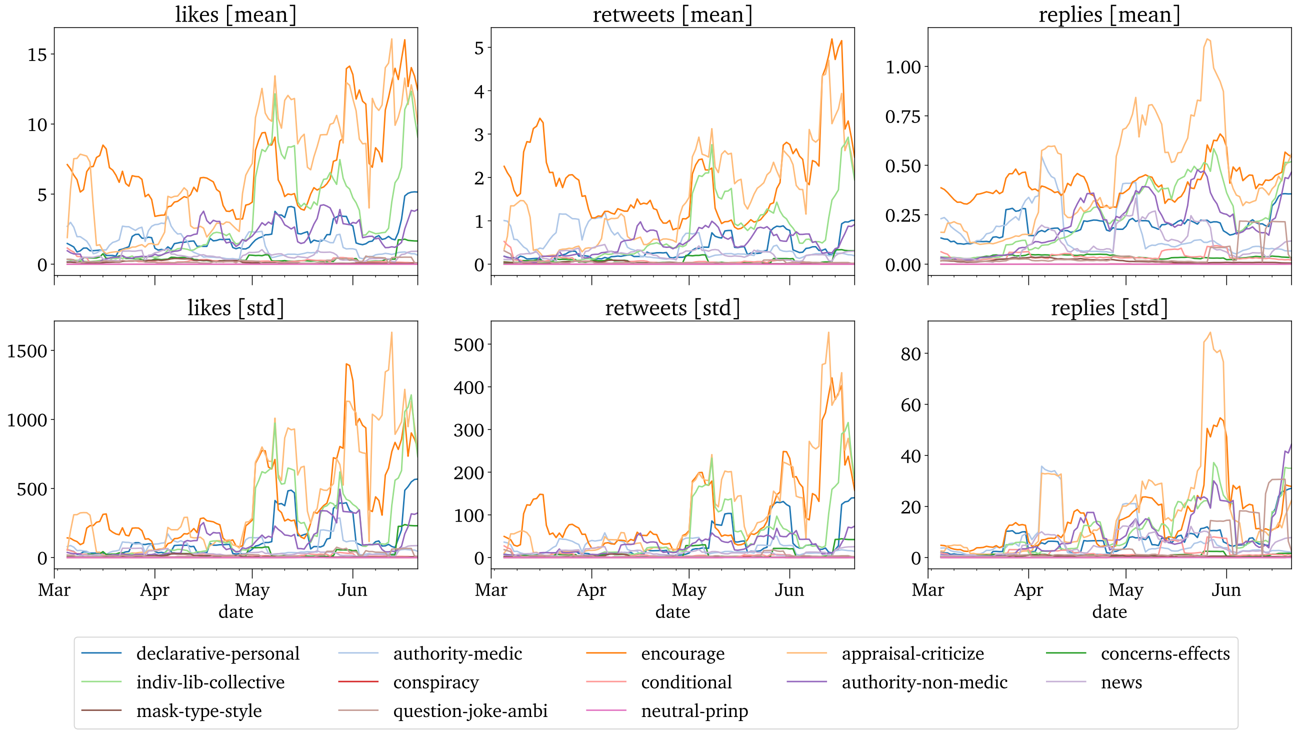


FIGURE 12. Upper row: seven-day rolling averages of the counts for likes, retweets, and replies received by tweets with different themes. Lower row: standard deviations of the averages.

all replies are antagonistic towards the original tweet, but non-antagonistic replies tend to be accompanied with likes, thus keeping replies from outnumbering likes.

The other two ratios, $R_{\text{retweets-likes}}$ and $R_{\text{replies-retweets}}$, do not have commonly accepted interpretations. Describing these ratios in terms of their relationship with $R_{\text{replies-likes}}$ and with each other can provide some illumination on what they could potentially represent. $R_{\text{replies-likes}}$ has a positive correlation (0.7093) with $R_{\text{replies-retweets}}$ and non-existent correlation (0.05) with $R_{\text{retweets-likes}}$. A negative correlation

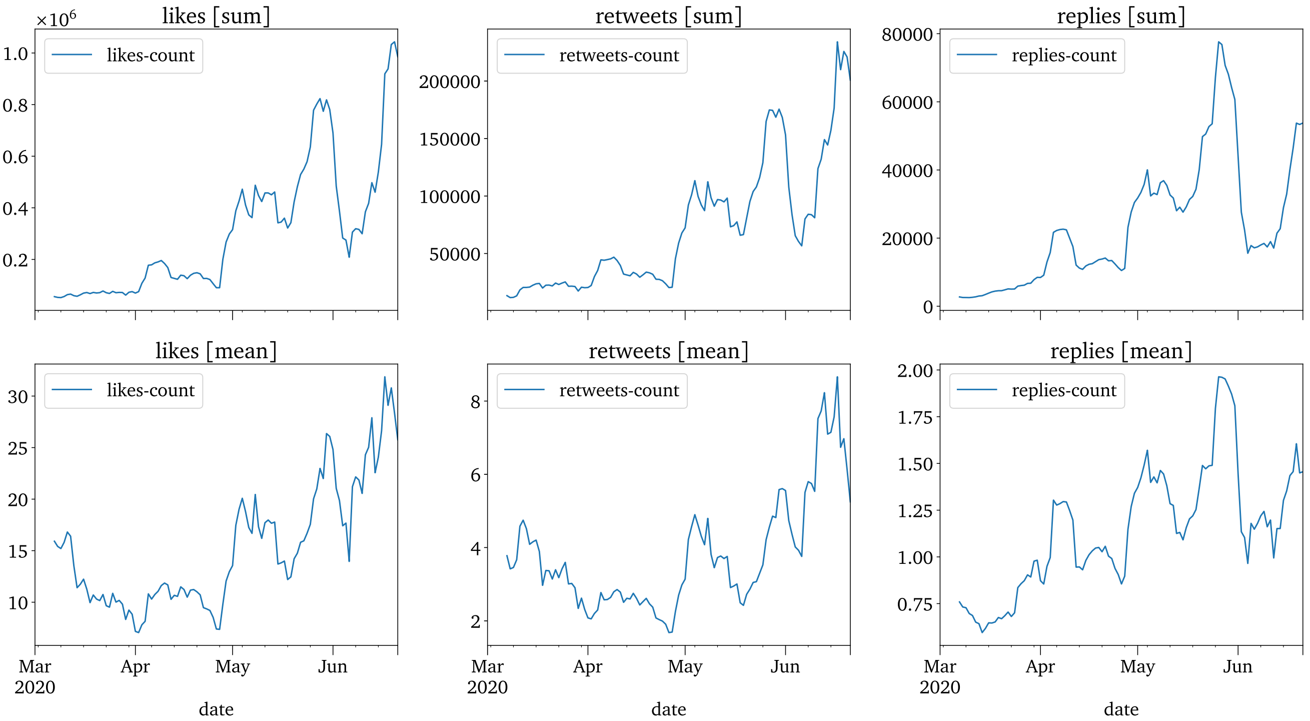


FIGURE 13. Seven-day rolling averages of the total counts (upper row) and mean counts (lower row) of likes, retweets, and replies received by tweets.

(-0.7067) exist between $R_{\text{replies-retweets}}$ and $R_{\text{retweets-likes}}$. There are 494,964 tweets with positive $R_{\text{replies-retweets}}$ and 41,851 tweets with positive $R_{\text{retweets-likes}}$. Of the 221,214 tweets with positive $R_{\text{replies-likes}}$, 216,366 (97.80%) tweets also have positive $R_{\text{replies-retweets}}$ while only 4,827 (2.18%) have positive $R_{\text{retweets-likes}}$.

Based on these intersections and correlations, our point of view is that $R_{\text{replies-retweets}}$ serves to identify disagreeable tweets in the same manner as $R_{\text{replies-likes}}$ where positive values indicate disagreeableness but it is a laxer criterion that also ends up capturing tweets that are not as offensive as those that have positive $R_{\text{replies-likes}}$. Positive $R_{\text{retweets-likes}}$ identify tweets with debatable to mildly disagreeable positions — the positions are not so egregious that their replies outnumber likes, hence it having near-zero intersection with positive $R_{\text{replies-likes}}$. This interpretation of $R_{\text{retweets-likes}}$ also dovetails with the negative correlation that it has with $R_{\text{replies-retweets}}$. As tweets go from contentious to being actively offensive, replies are more likely to outnumber retweets, leading to increasingly positive $R_{\text{replies-retweets}}$, while retweets are less likely to outnumber likes (not due to more people liking the tweets but perhaps due to fewer people retweeting the tweets), leading to negative $R_{\text{retweets-likes}}$.

One extra factor that needs to be kept in mind when examining the ratios for themes is that many themes are much rarer compared to other themes and thus have very low engagement, as can be seen by their respective proportions in Figure 5. This translates to few tweets with likes that outnumber retweets or replies by very significant margins, resulting in ratios that stay close to zero. If we look at the standard deviations of ratios (lower row of Figure 11), the only themes that are in the same league in terms of the magnitude of variations are *encourage*, *appraisal-criticize*, *indiv-lib-collective*, *declarative-personal*, and *authority-non-medic*. Therefore, we will avoid comparing ratiometric trends for themes outside this group.

The factor mentioned above does not apply to stance ratios as all three stances have similarly high standard deviations (lower row of Figure 9).

b: RATIOMETRIC AND METRIC TRENDS: STANCES

From March until late June, the daily average $R_{\text{retweets-likes}}$ for neutral tweets stayed above that of positive tweets most of the time, which in turned stayed above that of negative tweets all the time (Figure 9 left). In other words, neutral tweets are likelier to be merely controversial or slightly offensive. This could be because neutral tweets, by not committing to a position, are likelier to engender debate among its audience but are less capable of immediately making the audience dislike them compared to negative tweets. For the daily average $R_{\text{replies-retweets}}$ and $R_{\text{replies-likes}}$, the stances ranked in descending order goes negative, neutral, and positive (Figure 9 top center and right); negative tweets are more likely to be considered to be disagreeable than neutral or negative tweets.

From early April to late May, $R_{\text{retweets-likes}}$ for tweets with positive stances was on a downward trajectory, becoming almost as uncontroversial as tweets with negative stances. This does not automatically mean that people agree more with positive tweets. In the same period of time, looking at the $R_{\text{replies-retweets}}$ plot, we see the values for positive tweets going up, implying that positive tweets might be transitioning from being uncontroversial to being disagreeable. The June anomaly briefly reversed those trends, making positive tweets controversial again but much less likely to be disagreeable. Post-June anomaly saw another trend reversal for positive tweets but some gains were kept; positive tweets are once again less controversial and more disagreeable but it is not less controversial than negative tweets and the growth in disagreeableness only brought positive tweets back to the same level as the minimum in early April and not the peak in late May. The trend for extreme offensiveness ($R_{\text{replies-likes}}$) never saw significant changes for positive tweets, declining from March until late June.

The disagreeableness ($R_{\text{replies-retweets}}$ and $R_{\text{replies-likes}}$) for negative tweets was on a very slight down trend from early April to late May. The June anomaly period itself did not alter that trend. However, post-June anomaly, disagreeableness shot up. As for contentiousness ($R_{\text{retweets-likes}}$), it was on very slight upward trend for negative tweets from mid-April to late May, increased sharply during the June anomaly, then declined sharply post-June anomaly. The decline in contentiousness coincided with the increase in disagreeableness.

Around the start of April, due to Trump ignoring his administration's own mask mandate (Section IV-D3), disagreeableness for tweets of all stances increased briefly before decreasing sharply (Figure 9 top center and right). The influx of users (Figure 6) at first likely caused further polarization, causing the initial increase in disagreeableness. But as the incoming users were increasingly dispersed among many conversations instead of being concentrated in a few, which is not conducive to creating a situation where a small number of tweets can receive the majority of attention and replies, disagreeableness decreased.

The daily averages for likes, retweets, and replies (Figure 10) shows that positive tweets receive more likes and retweets than tweets of other stances from March through June. Between neutral and negative tweets, neutral tweets have a slight advantage in numbers. As for replies, neutral tweets receive the most on average while positive and neutral tweets have comparable figures, perhaps a reflection of neutral tweets' status as a place for the debate and exchange of viewpoints. In the days leading up to the June anomaly, we can see that a small number of negative tweets have received more likes and retweets than is usual for that time period, as we can see from the spikes in standard deviations (Figure 10 lower row). During the same time period, a small number of positive tweets have also received a very high amount of replies, which is evident in both the mean and standard deviation plots (Figure 10 right column).

c: RATIOMETRIC AND METRIC TRENDS: THEMES

To compare the engagement received by tweets with different themes, we plotted their ratiometrics in Figure 11 and the mean counts of their likes, retweets, and replies in Figure 12. We focus only on the *encourage*, *appraisal-criticize*, *indiv-lib-collective*, *declarative-personal*, and *authority-non-medic* themes due to the reason laid out near the end of Section IV-D5.a.

From early March to late May, tweets with an *indiv-lib-collective* or an *appraisal-criticize* theme became less and less controversial ($R_{\text{retweets-likes}}$). The June anomaly caused these two themes to very suddenly become more controversial, although post-June anomaly the contentiousness of the two themes once again trended downwards. In contrast, *encourage* tweets became more controversial from March until the end of April, then stayed the same level of contentiousness throughout May. June anomaly brought about a downward trend for *encourage* tweets' contentiousness that persisted until late June. *Declarative-personal* and *authority-non-medic* tweets became less controversial from March to the mid-April then stayed mostly level after that, remaining mostly unaffected by the June anomaly.

In terms of disagreeableness, tweets with a *indiv-lib-collective* or a *appraisal-criticize* theme have similar trends. More of them are likely to be found mildly offensive ($R_{\text{replies-retweets}}$) as time progresses from March until late May. However, tweets with those two themes are less and less likely to be found very offensive ($R_{\text{replies-likes}}$) over the same time window. During the June anomaly, those trends reversed briefly, i.e. less likely to be mildly offensive but more likely to be very offensive, before reverting again post-June anomaly, i.e. more likely to be mildly offensive but less likely to be very offensive. However, they did not reach a level as low as before the June anomaly. For *encourage* tweets, they became less likely to be mildly offensive from March to mid-April and less likely to very offensive from from early to late March. After those two points in time, *encourage* tweets became more likely to be both mildly offensive and very offensive. During the June anomaly, *encourage* tweets suddenly became much less likely to be mildly offensive and severely offensive. Post-June anomaly, *encourage* tweets stayed less likely to be severely offensive although they once again become likelier to be mildly offensive. The *declarative-personal* and *authority-non-medic* themes became more likely to be mildly offensive and less likely to be very offensive from March until late April. For the remaining rest of the time, there were some fluctuations on their level of offensiveness but none of the changes were as severe as that experienced by the other themes we have previously discussed.

The *indiv-lib-collective* and *appraisal-criticize* themes having the highest co-occurrence with each other (Figure 1) may explain why their ratio trends are so tightly intertwined. Co-occurrence does not offer a sufficient explanation for the similarity in trends seen in *declarative-personal* and *authority-non-medic*, as they do not have the highest co-occurrence with each other. For *declarative-personal*, it is highest with *indiv-lib-collective*. For *authority-non-medic*, it is highest with *appraisal-criticize* and *indiv-lib-collective*.

The mean counts of likes and retweets of *encourage* and *appraisal-criticize* tweets have been comparable since early April. The replies received by the two themes, however, are not. *Appraisal-criticize* tweets are prone to sudden surges of replies and they received far more replies than *encourage* tweets during the entire month of May. *Encourage* tweets, in contrast, received a consistent number of replies from March till June. In the days leading up to the June anomaly, a small number of tweets with either *appraisal-criticize* or *encourage* tweets did receive an extremely large number of replies compared to other tweets bearing the same themes, the result of which can be seen on the standard deviations of replies for the two themes (Figure 12 lower right).

6) BART SUMMMARIES OF THE FIVE LARGEST CONVERSATIONS IN LATE MAY

Out of the ten largest conversations found in our dataset (Table 2), five of them occurred in late May, during or just before the June anomaly, four in mid-June, near the end of our data collection period, and one in mid April. BART was used to summarize the five May conversations that could provide us with clues as to what might have potentially led to the June anomaly.

The starting dates (Dates (min.)) and summaries of the conversations, listed in the same order as Table 2, with spelling, punctuation, grammatical, and other errors preserved intact:

- 2020-05-25: The only reason Trump doesn't wear a mask is he has no leadership qualities. He doesn't wear a masks in public because he's a dumbass. And he could never out-cool President Biden. He looks like a responsible, compassionate human being. If you care about those around you, you'd worn a mask too.
- 2020-05-26: "Wearing a mask is useless if you don't wear it correctly! If you don't wear a mask, you ain't white. You're entitled to stay home a wear a masks. You can take your mask and shove it where the sun don't shine," she said.
- 2020-05-24: Wearing a mask at the grocery store isn't Nazi Germany…If you think a mask will protect you feel free to wear one. I'm not the dum dum going out coughing my lungs out on other people. If you refuse to wear a mask, don't go into a public space.
- 2020-05-29: "If you honestly believe a mask will save you, wear it. If that makes me not nice in some peoples minds, I can live with that" "I wear a mask when I go to the store, but I will admit I lift it numerous times to get air" "The anxiety when wearing one is much worse than my fear of the virus & I I'm in the high risk category"
- 2020-05-23: I don't give a shit if you wear a mask anymore than ifYou wear a t-shirt or a sweater. How is wearing a mask make one suffer? Just wear it. If you're going to wear a masks still, while driving your car, alone. You need to just stay home, you don""t deserve to be out. I'm hoping that knitted mask (that has holes) has a filter in between - otherwise,

From these five summaries, the situation of the mask-wearing discourse on Twitter just at the start of the June anomaly aligns with the stance proportions in Figure 5, where approximately three-tenths of the tweets in late May are anti-mask. The conversations dated 2020-05-26, 2020-05-24, 2020-05-29, and 2020-05-23 contain sentences that are anti-mask, some definitive (e.g., "You can take your mask and shove it wherethe sun don't shine") and some ambiguous (e.g., "If you honestly believe a mask will save you, wear it."). The conversation dated 2020-05-23 appears to concern Alyssa Milano wearing a crocheted mask in a car.[12] While participants of the Alyssa Milano conversation may be sincerely advocating for wearing better masks, some among them might be hijacking it to push an anti-mask agenda, e.g., telling people who voluntarily wear masks in cars to stay at home instead, hence the low mean stance. The largest conversation, with a start date of 2020-05-25, contained references to Trump and Biden, demonstrating that the politicized debates on mask-wearing attract far more attention than those without.

The presence of these five May conversations challenges the idea that the sharp drop in tweet volume during the June anomaly can simply be explained away as just user interest naturally falling off. First, to have five of the ten largest conversation in the entire dataset be located within a small time window in late May suggests that interest in tweeting about masking remains very strong at the beginning of the June anomaly, and if interest were to decay, it is likelier to decay slowly and be more reminiscent of what was seen in at the start of April. Second, the sentences in the summaries are a mix of positive, neutral, and negative in terms of their stances towards masking, suggesting that Twitter users have not all converged upon a single stance on the issue of mask-wearing and thus have lost interest in debating it. Third, going by the dates of the latest tweets to be posted within those five May conversations in Table 2, we can see that interest in them persisted until early and mid-June.

a: MONOLITHIC CONVERSATION

The second last conversation with a start date of 2020-06-16 (conversation ID: 1272953220520988673), which has the highest mean stance out of the ten listed at 4.86 out of a maximum of 5, serves as a good example of a monolithic conversation where tweets exhibiting a single stance overwhelmingly dominated. The conversation consists of people comparing the triviality of wearing masks to the true hardships suffered by generations past. The BART summary:

- 2020-06-16: My father survived Pearl Harbor, Iwo Jima, Guadalcanal & severe malaria. My uncle survived the Bataan Death March. My grandfather had to hide under his dead soldier brothers in the Pacific Ocean campaign. My grandmother plowed fields in her bare feet. I am proud to wear a mask.

V. JUNE ANOMALY: CENSORSHIP?

A. TWITTER'S HISTORY OF CENSORSHIP

That Twitter engages in censorship, whether crudely through outright removal of tweets or more skillfully by reducing a tweet's visibility or by coloring user perception of a tweet's content, is not some outlandish supposition.

[12] https://news.yahoo.com/alyssa-milano-defends-totally-safe-230603836.html

TABLE 2. **The IDs of the ten largest conversations found in our dataset along with the earliest and latest dates of tweets associated with the conversation found in the dataset; the counts of tweets associated with each conversation; the total number of users, likes, retweets, and replies; the mean stance; and the top two themes and their respective counts.**

Conversation ID	Date (min.)	Date (max.)	# tweets	# users	# likes	# retweets	# replies	stance	themes (theme counts)
1265045009323241472	2020-05-25	2020-06-20	4,162	3824	117,733	21,965	73,688	4.08	*appraisal-criticize* (3,584) *indiv-lib-collective* (1,460)
1265445256465731586	2020-05-26	2020-06-15	2,456	2240	523,654	75,573	41,893	2.81	*appraisal-criticize* (1,650) *encourage* (972)
1264721468761481216	2020-05-24	2020-06-13	2,260	1728	20,070	828	2,583	3.15	*indiv-lib-collective* (1,513) *appraisal-criticize* (1,258)
1266406045544579073	2020-05-29	2020-06-06	2,182	2022	8,028	1,062	1,160	2.39	*indiv-lib-collective* (1,586) *appraisal-criticize* (1,084)
1273295136429084672	2020-06-17	2020-06-21	2,171	1779	37,045	12,471	9,353	3.07	*indiv-lib-collective* (1,387) *encourage* (1,077)
1271966131230699520	2020-06-13	2020-06-18	2,164	2111	70,991	3,366	26,623	4.64	*declarative-personal* (1,773) *indiv-lib-collective* (459)
1250130325356888064	2020-04-14	2020-06-01	2,126	1998	4,454	222	547	3.89	*encourage* (1,421) *appraisal-criticize* (1,136)
1272953220520988673	2020-06-16	2020-06-20	2,032	1975	180,521	20,341	3,468	4.86	*declarative-personal* (1,944) *indiv-lib-collective* (331)
1264248352683646976	2020-05-23	2020-06-09	1,758	1507	11,345	780	2,188	2.44	*appraisal-criticize* (1,055) *indiv-lib-collective* (787)
1273018665391099904	2020-06-16	2020-06-20	1,664	1515	25,756	1,942	2,313	3.88	*indiv-lib-collective* (1,242) *appraisal-criticize* (1,208)

While the process is opaque, Twitter's heavy-handed moderation tactics is a publicly known fact. Banning and removing tweets are not the only tools at the moderation team's disposal. Twitter has clarified that while it does not engage in shadow banning,[13] it does deliberately and severely reduce the visibility of tweets from "bad faith actors" [88], e.g., search result rank manipulation to bury targeted tweets, essentially consigning tweets to death through near-obscurity. As it pertains to the COVID-19 pandemic, Twitter has implemented a new company policy on disinformation and removed tweets they deemed harmful from highly visible world leaders [89] back in April of 2020, which is within our data's date range. And more recently in April 2021, well outside our data date range, Twitter has removed tweets critical of the Indian government's coronavirus response at the behest of the Indian government [90].

There has been research on the plausibility of shadow banning on Twitter [91], with the authors finding that visibility limitations on user profiles being bugs, as claimed by Twitter, "is statistically unlikely with regards to the data [the authors] collected". Twitter has also attempted to steer certain conversations through the use of soft moderation tools like the very recent policy of adding warning labels to tweets concerning presidential election results and COVID-19. This new warning label policy has also garnered the attention of researchers [92], who found that among tweets stuck with the label, 72% were shared by Republicans and 11% by Democrats, perhaps demonstrating a political bias. Researchers have also suggested that Twitter's failure to communicate the reasons for censoring content has led to many users believing that censorship is politically motivated [93].

B. EVIDENCE FROM OUR DATASET

Throughout Section IV-D, the methods we used to probe our dataset have provided simple explanations for the causes behind temporal trends found in our dataset. The only exception is the June anomaly, where instead they have served to successively eliminate simpler explanations until censorship became the simplest remaining explanation. We collate our findings here.

Aside from the anomalous drop in tweet count, user interest in mask-wearing as a topic did not show signs of abating during the June anomaly. A heated debate was still occurring prior to the June anomaly, judging by the high number of replies that tweets were receiving in late May (Figure 13 and Sections IV-D5.b and IV-D5.c). Five out of the ten largest conversations in our dataset started around late May and the latest tweets within those conversations can be found after the June anomaly, demonstrating the there is continuity in user interest (Section IV-D6). The swift recovery in tweet count post-June anomaly is another sign that user interest did not truly decline (Section IV-D3). The June anomaly is therefore unlikely to be a product of user interest suddenly dropping off.

The unequal impact that the June anomaly has had on tweets with different stances and themes is further evidence that the anomaly was an engineered event. Just before the June anomaly, a small number of negative tweets were being boosted by a very high number of likes and retweets while a small number of positive retweets were assaulted with a high number of replies (Section IV-D5.b). The share of negative tweets was also growing day by day prior to the June anomaly (Section IV-D2). Everything changed with the June anomaly. Not only did the June anomaly increase the proportion of positive tweets, these tweets were also of a good quality, i.e. less likely to be severely disagreeable (Section IV-D5.b). In contrast, fewer negative tweets were

[13] "[D]eliberately making someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster" [88].

found and these tweets are of low quality. Mean stances for the *encourage*, *authority-non-medic*, and *authority-medic* themes reversed their declines. The proportion of *encourage*-themed tweets, which are strongly associated with a positive stance, increased while the contentiousness and disagreeableness for *encourage* tweets decreased. Proportions for *appraisal-criticize* and *indiv-lib-collective*, which have more equal associations and thus can be considered as more ''argumentative'', were reduced. In essence, users are more likely to encounter pro-mask tweets and less likely to be aware of dissenting opinions or even the existence of dissenting opinions.

VI. DISCUSSION

Whether our argument that the June anomaly was an act of censorship has been sufficiently persuasive or not, we believe that any study involving Twitter data should dispense with the unspoken assumption that Twitter is a neutral platform that is only at the mercy of malicious outside actors. Internal actors matter. Twitter's moderation exerting an influence over the evolution of the conversations should be accounted for. From attaching warning labels to content shared mainly by Republicans to banning tweets critical of the Indian government's COVID-19 response, Twitter has clearly never aimed to have the fairest representation of opinions. Results obtained from Twitter data is not just unrepresentative of the general public due to Twitter's demographics, it is also potentially unrepresentative of Twitter's own users as certain opinions may be suppressed.

Aside from identifying potential censorship, we also believe that the results of our analyses can offer clues as to how one can best affect change in the overall stance of a contentious issue on Twitter, at least within the narrow niche of mask-wearing for the early months of a coronavirus pandemic to a US-centric, if not necessarily American, audience. These clues can be useful because a lot of policies cannot be unilaterally imposed top-down in liberal democracies or at least in societies that wish to maintain a facade of individual liberty taking precedence over every other priority — the public has to voluntarily accept them. Take the COVID-19 pandemic for example. Successful management of the pandemic requires managing the spread of undesirable information that galvanizes resistance against intervention methods such as vaccinations and universal masking.

As to the question of ethicality of the suggestions we offer here, it is beyond the scope of our work. The debate of whether lofty libertarian ideals encapsulated by quotes such as ''I disapprove of what you say, but I will defend to the death your right to say it'' and ''those who would give up essential Liberty, to purchase a little temporary Safety, deserve neither Liberty nor Safety'', or if pragmatic sacrifices need to be made for the greater good in exceptional circumstances through overt censorship or more covert ones such as tarring and feathering dissent as disinformation and misinformation, is a debate that would not be resolved in this paper.

From the perspective of people who have direct or indirect control over social media platforms such as the platform owners themselves or the government under whose auspices and jurisdiction the platform operates, an effective short-term tactic for managing conversations that are trending in an undesirable direction is to minimize all discussion, even if it meant temporarily silencing supportive voices (during the June anomaly, the absolute tweet count for positive tweets went down). While the Streisand effect does not apply within the platform due to having control over what messages are visible on the platform, people may move to a viable alternative social media platform (e.g., the platform has an audience of comparable size) to escape censorship. If no viable alternative exist, users may resort to circumventing moderation, e.g., discussing a topic in an oblique manner to escape word filters, provided that the users are aware that discussion is being suppressed in the first place.

From the perspective of those without moderation powers over a platform, efforts to influence the prevailing stance in a conversation or on the entire platform is better done by focusing all efforts on a few conversations with larger audiences than diluting those efforts over many conversations. This is based on our comparison of two events where a large number of users began participating in mask-wearing conversations, one in March and another in April, in which the one where the new users congregated on a few conversations altered the mean stance while the other where the new users are dispersed did not have a visible impact. The threat of centralized attention may have led to the June anomaly; in the days leading up to the June anomaly, a small number of positive and *appraisal-criticize* tweets were receiving a very large number of replies, based on the high means and high standard deviations at the time. If there are no existing conversations with a large number of participants to hijack, one can attempt to build an audience by politicizing an issue. This is based on the observation that the conversation with the largest number of participants in our dataset being more about Trump and Biden than mask-wearing itself. An influence operation can also involve supplanting a narrative with a counter-narrative that has the same themes. This insight was gleaned from observing the ''store employees not being allowed to wear mask'' narrative shift to ''stores making customers wear mask is oppression'' with the resultant drop in mean stance for those themes.

VII. FUTURE WORK

While American influence is far-reaching and can impact lockdown protests in Germany[14] and anti-mask movements in Quebec,[15] the findings in this paper is still based upon a primarily English dataset that represents US-centric viewpoints. The frequency of tweets with anti-mask stance following a Pacific Daylight Time diurnal pattern and the predominance of US politicians and celebrities as conversation topics demonstrate the limitations facing any attempts at generalizing the findings from this paper. The social media platform we have chosen, Twitter, which has moderation policies as well

[14] https://slate.com/technology/2020/09/qanon-europe-germany-lockdown-protests.html

[15] https://www.cbc.ca/news/canada/montreal/quebec-anti-mask-movement-qanon-covid-19-1.5737040

other platform-specific features that encourages its users to interact with it in a certain manner (e.g., ratioing tweets) also limits the generalizabiilty of our findings.

Wearing a mask to combat an airborne pandemic is novel in that until the COVID-19 pandemic, most people in the Anglosphere have never had to grapple with adhering to such a policy. Therefore, people might be more susceptible to being dislodged from their present stance on masking. Whether influence can be as easily gained or lost through social media manipulation for topics where the proponents and detractors are more entrenched, e.g., vaccine hesitancy, remains to be studied.

Due to all of the issues we have listed above, we believe that future work investigating the same topic of mask-wearing conversations could explore other languages, countries, and social media platforms so that we can determine if commonalities exist among these different contexts.

VIII. CONCLUSION

Studying how influence is won or lost on the contentious healthcare issue of mask-wearing on a widely-used social media platform, Twitter, provided helpful clues on designing information interventions that can help interested parties in more effectively controlling the narrative of such issues online. Stance and theme labels on tweets helped identify inflection points where opinions on the issue changed. Used in conjunction with Twitter's metrics, the labels also provided insights on the factors driving the evolution of the mask-wearing discourse; concentrating attention and breaking concentrated attention appears to be the optimal strategy for propagating and halting the propagation of stances on Twitter.

Using stances and themes to probe a certain anomalous event in our dataset has also highlighted the limitations of using Twitter data as a way to accurately poll public opinion on controversial issues and the need to discard the assumption that the platform allows all opinions to be represented equally.

Our hope is that our findings here will help in crafting new information manipulation strategies that can be tested for generalizability on other issues, from more closely related topics and social media platforms such as propagandizing the wearing of masks for seasonal flus on Twitter to more distant ones such as broadening climate change acceptance on Facebook.

COMPETING INTERESTS STATEMENT

The authors have no competing interests to declare.

REFERENCES

[1] J. Steere-Williams, "Endemic fatalism and why it will not resolve COVID-19," *Public Health*, vol. 206, pp. 29–30, May 2022.

[2] A. C. K. Lee and J. R. Morling, "Living with endemic COVID-19," *Public Health*, vol. 205, pp. 26–27, Apr. 2022.

[3] S. Carazo, D. M. Skowronski, M. Brisson, C. Sauvageau, N. Brousseau, R. Gilca, M. Ouakki, S. Barkati, J. Fafard, D. Talbot, V. Gilca, G. Deceuninck, C. Garenc, A. Carignan, P. De Wals, and G. De Serres, "Estimated protection of prior SARS-CoV-2 infection against reinfection with the Omicron variant among messenger RNA-vaccinated and nonvaccinated individuals in Quebec, Canada," *JAMA Netw. Open*, vol. 5, no. 10, Oct. 2022, Art. no. e2236670.

[4] A. S. Breathnach, P. A. Riley, M. P. Cotter, A. C. Houston, M. S. Habibi, and T. D. Planche, "Prior COVID-19 significantly reduces the risk of subsequent infection, but reinfections are seen after eight months," *J. Infection*, vol. 82, no. 4, pp. e11–e12, Apr. 2021.

[5] J. L. Bernal, N. Andrews, C. Gower, E. Gallagher, R. Simmons, S. Thelwall, J. Stowe, E. Tessier, N. Groves, G. Dabrera, R. Myers, C. N. J. Campbell, G. Amirthalingam, M. Edmunds, M. Zambon, K. E. Brown, S. Hopkins, M. Chand, and M. Ramsay, "Effectiveness of COVID-19 vaccines against the B.1.617.2 (delta) variant," *New England J. Med.*, vol. 385, no. 7, pp. 585–594, Aug. 2021.

[6] E. Mahase, "COVID-19: Novavax vaccine efficacy is 86% against U.K. variant and 60% against South African variant," *BMJ*, vol. 372, p. n296, Feb. 2021.

[7] D. Zhou et al., "Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera," *Cell*, vol. 184, no. 9, pp. 2348–2361, Feb. 2021.

[8] E. Eythorsson, H. L. Runolfsdottir, R. F. Ingvarsson, M. I. Sigurdsson, and R. Palsson, "Rate of SARS-CoV-2 reinfection during an Omicron wave in Iceland," *JAMA Netw. Open*, vol. 5, no. 8, Aug. 2022, Art. no. e2225320.

[9] J. R. C. Pulliam, C. van Schalkwyk, N. Govender, A. von Gottberg, C. Cohen, M. J. Groome, J. Dushoff, K. Mlisana, and H. Moultrie, "Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa," *Science*, vol. 376, no. 6593, May 2022, Art. no. eabn4947.

[10] R. Grewal, L. Nguyen, S. A. Buchan, S. E. Wilson, S. Nasreen, P. C. Austin, K. A. Brown, D. B. Fell, J. B. Gubbay, K. L. Schwartz, M. Tadrous, K. Wilson, and J. C. Kwong, "Effectiveness of mRNA COVID-19 vaccine booster doses against Omicron severe outcomes," *Nature Commun.*, vol. 14, no. 1, p. 1273, Mar. 2023.

[11] J. J. Lau, S. M. S. Cheng, K. Leung, C. K. Lee, A. Hachim, L. C. H. Tsang, K. W. H. Yam, S. Chaothai, K. K. H. Kwan, Z. Y. H. Chai, T. H. K. Lo, M. Mori, C. Wu, S. A. Valkenburg, G. K. Amarasinghe, E. H. Y. Lau, D. S. C. Hui, G. M. Leung, M. Peiris, and J. T. Wu, "Real-world COVID-19 vaccine effectiveness against the Omicron BA.2 variant in a SARS-CoV-2 infection-naive population," *Nature Med.*, vol. 29, no. 2, pp. 348–357, Feb. 2023.

[12] S. Chalkias, C. Harper, K. Vrbicky, S. R. Walsh, B. Essink, A. Brosz, N. McGhee, J. E. Tomassini, X. Chen, Y. Chang, A. Sutherland, D. C. Montefiori, B. Girard, D. K. Edwards, J. Feng, H. Zhou, L. R. Baden, J. M. Miller, and R. Das, "A bivalent Omicron-containing booster vaccine against COVID-19," *New England J. Med.*, vol. 387, no. 14, pp. 1279–1291, Oct. 2022.

[13] S. M. Scheaffer et al., "Bivalent SARS-CoV-2 mRNA vaccines increase breadth of neutralization and protect against the BA.5 Omicron variant in mice," *Nature Med.*, vol. 29, no. 1, pp. 247–257, Jan. 2023.

[14] I. González-Domínguez, J. L. Martínez, S. Slamanig, N. Lemus, Y. Liu, T. Y. Lai, J. M. Carreño, G. Singh, G. Singh, M. Schotsaert, I. Mena, S. McCroskery, L. Coughlan, F. Krammer, A. García-Sastre, P. Palese, and W. Sun, "Trivalent NDV-HXP-S vaccine protects against phylogenetically distant SARS-CoV-2 variants of concern in mice," *Microbiol. Spectr.*, vol. 10, no. 3, Jun. 2022, Art. no. e0153822.

[15] K. A. Parham, G. N. Kim, C. G. Richer, M. Ninkov, K. Wu, N. Saeedian, Y. Li, R. Rashu, S. D. Barr, E. J. Arts, S. M. M. Haeryfar, C. Y. Kang, and R. M. Troyer, "Monovalent and trivalent VSV-based COVID-19 vaccines elicit neutralizing antibodies and CD8$^+$ T cells against SARS-CoV-2 variants," *iScience*, vol. 26, no. 4, Apr. 2023, Art. no. 106292.

[16] J. Aw, J. J. B. Seng, S. S. Y. Seah, and L. L. Low, "COVID-19 vaccine hesitancy—A scoping review of literature in high-income countries," *Vaccines*, vol. 9, no. 8, p. 900, Aug. 2021.

[17] M. Sallam, "COVID-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates," *Vaccines*, vol. 9, no. 2, p. 160, Feb. 2021.

[18] P. N. Mutombo, M. P. Fallah, D. Munodawafa, A. Kabel, D. Houeto, T. Goronga, O. Mweemba, G. Balance, H. Onya, R. S. Kamba, M. Chipimo, J.-M.-N. Kayembe, and B. Akanmori, "COVID-19 vaccine hesitancy in Africa: A call to action," *Lancet Global Health*, vol. 10, no. 3, pp. e320–e321, Mar. 2022.

[19] O. J. Wouters, K. C. Shadlen, M. Salcher-Konrad, A. J. Pollard, H. J. Larson, Y. Teerawattananon, and M. Jit, "Challenges in ensuring global access to COVID-19 vaccines: Production, affordability, allocation, and deployment," *Lancet*, vol. 397, no. 10278, pp. 1023–1034, Mar. 2021.

[20] M. M. Kavanagh, L. O. Gostin, and M. Sunder, "Sharing technology and vaccine doses to address global vaccine inequity and end the COVID-19 pandemic," *JAMA*, vol. 326, no. 3, pp. 219–220, Jul. 2021.

[21] S. S. Bajaj, L. Maki, and F. C. Stanford, ''Vaccine apartheid: Global cooperation and equity,'' *Lancet*, vol. 399, no. 10334, pp. 1452–1453, Apr. 2022.

[22] U. A. Mejias and N. E. Vokuev, ''Disinformation and the media: The case of Russia and Ukraine,'' *Media, Culture Soc.*, vol. 39, no. 7, pp. 1027–1042, Oct. 2017.

[23] V. Carrieri, L. Madio, and F. Principe, ''Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy,'' *Health Econ.*, vol. 28, no. 11, pp. 1377–1382, Nov. 2019.

[24] E. Pertwee, C. Simas, and H. J. Larson, ''An epidemic of uncertainty: Rumors, conspiracy theories and vaccine hesitancy,'' *Nature Med.*, vol. 28, no. 3, pp. 456–459, Mar. 2022.

[25] A. Kohut, C. Doherty, M. Dimock, and S. Keeter, ''Further decline in credibility ratings for most news organizations,'' Pew Res. Center, Washington, DC, USA, Tech. Rep., Aug. 2012. [Online]. Available: https://www.pewresearch.org/politics/2012/08/16/further-decline-in-credibility-ratings-for-most-news-organizations/

[26] S. Kemp, ''WhatsApp is the world's favorite social platform (and other facts),'' Social Media Marketing Manag. Dashboard, Hootsuite, 2023. [Online]. Available: https://blog.hootsuite.com/simon-kemp-social-media/

[27] B. Auxier and M. Anderson, ''Social media use in 2021,'' Pew Res. Center, Washington, DC, USA, Tech. Rep., Apr. 2021. [Online]. Available: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

[28] A. Nazaryan. (Sep. 2020). *CDC Chief Says Masks Better at Stopping Coronavirus than a Vaccine*. Yahoo News. [Online]. Available: https://news.yahoo.com/cdc-chief-says-masks-better-at-stopping-coronavirus-than-a-vaccine-173526486.html

[29] K. Hernandez. (Sep. 2020). *I'm a Doctor and Here's Why Masks are Better Than Vaccines*. Yahoo Life. [Online]. Available: https://www.yahoo.com/lifestyle/im-doctor-heres-why-masks-145348019.html

[30] M. Gandhi and G. W. Rutherford, ''Facial masking for COVID-19—Potential for 'variolation' as we await a vaccine,'' *New England J. Med.*, vol. 383, no. 18, p. e101, Oct. 2020.

[31] J. E. Haberer, A. Straten, S. A. Safren, M. O. Johnson, K. R. Amico, C. Rio, M. Andrasik, I. B. Wilson, and J. M. Simoni, ''Individual health behaviours to combat the COVID-19 pandemic: Lessons from HIV socio-behavioural science,'' *J. Int. AIDS Soc.*, vol. 24, no. 8, Aug. 2021, Art. no. e25771.

[32] R. O. Valdiserri, D. R. Holtgrave, and S. C. Kalichman, ''Barrier methods for the prevention of infectious diseases: Decades of condom research can inform the promotion of face mask use,'' *AIDS Behav.*, vol. 24, no. 12, pp. 3283–3287, Dec. 2020.

[33] L. Cao and Q. Liu, ''COVID-19 modeling: A review,'' 2021, *arXiv:2104.12556*.

[34] V. Batzdorfer, H. Steinmetz, M. Biella, and M. Alizadeh, ''Conspiracy theories on Twitter: Emerging motifs and temporal dynamics during the COVID-19 pandemic,'' *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 315–333, Dec. 2021.

[35] M. Vlasceanu and A. Coman, ''The impact of information sources on COVID-19 knowledge accumulation and vaccination intention,'' *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 287–298, Jan. 2022.

[36] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch, and S. Dietze, ''TweetsCOV19—A knowledge base of semantically annotated tweets about the COVID-19 pandemic,'' in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2991–2998.

[37] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, ''The COVID-19 social media infodemic,'' *Sci. Rep.*, vol. 10, no. 1, p. 16598, Oct. 2020.

[38] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, ''An 'Infodemic': Leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak,'' *Open Forum Infectious Diseases*, vol. 7, Jul. 2020, Art. no. ofaa258.

[39] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, ''Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention,'' *Psychol. Sci.*, vol. 31, no. 7, pp. 770–780, Jun. 2020.

[40] S. Z. Akbar, A. Panda, D. Kukreti, A. Meena, and J. Pal, ''Misinformation as a window into prejudice: COVID-19 and the information environment in India,'' in *Proc. ACM Human-Comput. Interact.*, vol. 4, no. CSCW3, pp. 249:1–249:28, Jan. 2021.

[41] A. M. Jamison, D. A. Broniatowski, M. Dredze, A. Sangraula, M. C. Smith, and S. C. Quinn, ''Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 'infodemic' on Twitter,'' *Harvard Kennedy School Misinformation Rev.*, vol. 1, pp. 1–22, Sep. 2020.

[42] L. Ermakova, D. Nurbakova, and I. Ovchinnikova, ''Analysis of users engaged in online discussions about controversial COVID-19 treatments,'' in *Proc. Adjunct 29th ACM Conf. User Modeling, Adaptation Personalization*, Utrecht, The Netherlands, Jun. 2021, pp. 162–166.

[43] J. S.-L. Kwan and K. H. Lim, ''Understanding public sentiments, opinions and topics about COVID-19 using Twitter,'' in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Dec. 2020, pp. 623–626.

[44] X. Yu, C. Zhong, D. Li, and W. Xu, ''Sentiment analysis for news and social media in COVID-19,'' in *Proc. 6th ACM SIGSPATIAL Int. Workshop Emergency Manage. GIS*, Washington, DC, USA, Nov. 2020, pp. 1–4.

[45] Z. Lu, Y. Jiang, C. Shen, M. C. Jack, D. Wigdor, and M. Naaman, ''''Positive energy'': Perceptions and attitudes towards COVID-19 information on social media in China,'' in *Proc. ACM Human-Comput. Interact. (CSCW)*, Apr. 2021, vol. 5, no. 1, pp. 177:1–177:25.

[46] L. G. Malagoli, J. Stancioli, C. H. G. Ferreira, M. Vasconcelos, A. P. C. da Silva, and J. M. Almeida, ''A look into COVID-19 vaccination debate on Twitter,'' in *Proc. 13th ACM Web Sci. Conf.*, Jun. 2021, pp. 225–233.

[47] L.-A. Cotfas, C. Delcea, R. Gherai, and I. Roxin, ''Unmasking people's opinions behind mask-wearing during COVID-19 pandemic—A Twitter stance analysis,'' *Symmetry*, vol. 13, no. 11, p. 1995, Oct. 2021.

[48] A. C. Sanders, R. C. White, L. S. Severson, R. Ma, R. McQueen, H. C. Alcântara Paulo, Y. Zhang, J. S. Erickson, and K. P. Bennett, ''Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse,'' in *Proc. AMIA Summits Transl. Sci.*, May 2021, pp. 555–564.

[49] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, ''SemEval-2016 Task 6: Detecting stance in tweets,'' in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, San Diego, CA, USA, 2016, pp. 31–41.

[50] W. Ahmed, P. A. Bath, L. Sbaffi, and G. Demartini, ''Novel insights into views towards H1N1 during the 2009 pandemic: A thematic analysis of Twitter data,'' *Health Inf. Libraries J.*, vol. 36, no. 1, pp. 60–72, Mar. 2019.

[51] C. Chew and G. Eysenbach, ''Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak,'' *PLoS One*, vol. 5, no. 11, Nov. 2010, Art. no. e14118.

[52] A. Signorini, A. M. Segre, and P. M. Polgreen, ''The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic,'' *PLoS One*, vol. 6, no. 5, May 2011, Art. no. e19467.

[53] S. B. Meyer, S. K. Lu, L. Hoffman-Goetz, B. Smale, H. MacDougall, and A. R. Pearce, ''A content analysis of newspaper coverage of the seasonal flu vaccine in Ontario, Canada, October 2001 to March 2011,'' *J. Health Commun.*, vol. 21, no. 10, pp. 1088–1097, Oct. 2016.

[54] V. Marivate, A. Moodley, and A. Saba, ''Extracting and categorising the reactions to COVID-19 by the South African public—A social media study,'' 2020, *arXiv:2006.06336*.

[55] S. Park, S. Han, J. Kim, M. M. Molaie, H. D. Vu, K. Singh, J. Han, W. Lee, and M. Cha, ''COVID-19 discourse on Twitter in four Asian countries: Case study of risk communication,'' *J. Med. Internet Res.*, vol. 23, no. 3, Mar. 2021, Art. no. e23272.

[56] D. Freelon and T. Lokot, ''Russian Twitter disinformation campaigns reach across the American political spectrum,'' *Harvard Kennedy School Misinf. Rev.*, vol. 1, no. 1, pp. 1–9, Jan. 2020.

[57] E. Bonsón, D. Perea, and M. Bednárová, ''Twitter as a tool for citizen engagement: An empirical study of the Andalusian municipalities,'' *Government Inf. Quart.*, vol. 36, no. 3, pp. 480–489, Jul. 2019.

[58] X. Yuan, R. J. Schuchard, and A. T. Crooks, ''Examining emergent communities and social bots within the polarized online vaccination debate in Twitter,'' *Social Media Soc.*, vol. 5, no. 3, Jul. 2019, Art. no. 2056305119865465.

[59] S. Chen, D. Khashabi, W. Yin, C. Callison-Burch, and D. Roth, ''Seeing things from a different angle: Discovering diverse perspectives about claims,'' in *Proc. Conf. North*, 2019, pp. 542–557.

[60] C. R. MacIntyre and A. A. Chughtai, ''Facemasks for the prevention of infection in healthcare and community settings,'' *BMJ*, vol. 350, no. 9, p. h694, Apr. 2015.

[61] D. K. Chu et al., ''Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: A systematic review and meta-analysis,'' *Lancet*, vol. 395, no. 10242, pp. 1973–1987, Jun. 2020.

[62] M. Liang, L. Gao, C. Cheng, Q. Zhou, J. P. Uy, K. Heiner, and C. Sun, ''Efficacy of face mask in preventing respiratory virus transmission: A systematic review and meta-analysis,'' *Travel Med. Infectious Disease*, vol. 36, Jul. 2020, Art. no. 101751.

[63] C. R. MacIntyre and A. A. Chughtai, "A rapid systematic review of the efficacy of face masks and respirators against coronaviruses and other respiratory transmissible viruses for the community, healthcare workers and sick patients," *Int. J. Nursing Stud.*, vol. 108, Aug. 2020, Art. no. 103629.

[64] J. T. Brooks and J. C. Butler, "Effectiveness of mask wearing to control community spread of SARS-CoV-2," *JAMA*, vol. 325, no. 10, pp. 998–999, Mar. 2021.

[65] V. C.-C. Cheng, S.-C. Wong, V. W.-M. Chuang, S. Y.-C. So, J. H.-K. Chen, S. Sridhar, K. K.-W. To, J. F.-W. Chan, I. F.-N. Hung, P.-L. Ho, and K.-Y. Yuen, "The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2," *J. Infection*, vol. 81, no. 1, pp. 107–114, Jul. 2020.

[66] W. Lyu and G. L. Wehby, "Community use of face masks and COVID-19: Evidence from a natural experiment of state mandates in the U.S.," *Health Affairs*, vol. 39, no. 8, pp. 1419–1425, Aug. 2020.

[67] A. Bałazy, M. Toivola, A. Adhikari, S. K. Sivasubramani, T. Reponen, and S. A. Grinshpun, "Do N95 respirators provide 95% protection level against airborne viruses, and how adequate are surgical masks?" *Amer. J. Infection Control*, vol. 34, no. 2, pp. 51–57, Mar. 2006.

[68] A. Konda, A. Prakash, G. A. Moss, M. Schmoldt, G. D. Grant, and S. Guha, "Aerosol filtration efficiency of common fabrics used in respiratory cloth masks," *ACS Nano*, vol. 14, no. 5, pp. 6339–6347, May 2020.

[69] E. Mahase, "COVID-19: What is the evidence for cloth masks?" *BMJ*, vol. 369, p. m1422, Apr. 2020.

[70] *Advice on the Use of Masks in the Context of COVID-19: Interim Guidance, 6 April 2020*, document WHO/2019-nCov/IPC_Masks/2020.3, World Health Org., Geneva, Switzerland, 2020.

[71] *Advice on the Use of Masks in the Context of COVID-19: Interim Guidance, 5 June 2020*, document WHO/2019-nCov/IPC_Masks/2020.4, World Health Org., Geneva, Switzerland, 2020.

[72] S. R. Kelleher. (May 2020). *Why Some Cities and Counties are Banning Face Masks with Valves*. Forbes. [Online]. Available: https://www.forbes.com/sites/suzannerowankelleher/2020/05/26/why-some-cities-and-counties-are-banning-face-masks-with-valves/

[73] L. Portnoff, J. Schall, J. Brannen, N. Suhon, K. Strickland, and J. Meyers, "Filtering facepiece respirators with an exhalation valve: Measurements of filtration efficiency to evaluate their potential for source control," U.S. Dept. Health Human Services, Public Health Service, Centers Disease Control Prevention, Nat. Inst. Occupational Safety Health, Washington, DC, USA, DHHS (NIOSH) Publication, Tech. Rep. 2021-107, Dec. 2020.

[74] M. S. Linneberg and S. Korsgaard, "Coding qualitative data: A synthesis guiding the novice," *Qualitative Res. J.*, vol. 19, no. 3, pp. 259–270, Jul. 2019.

[75] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "STANCY: Stance classification based on consistency cues," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 6412–6417.

[76] S. Ollinger, L. Dumani, P. Sahitaj, R. Bergmann, and R. Schenkel, "Same side stance classification task: Facilitating argument stance classification by fine-tuning a BERT model," 2020, *arXiv:2004.11163*.

[77] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Trans. Internet Technol.*, vol. 17, no. 3, pp. 1–23, Aug. 2017.

[78] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[79] A. Antelmi, D. Malandrino, and V. Scarano, "Characterizing the behavioral evolution of Twitter users and the truth behind the 90-9-1 rule," in *Proc. Companion Proc. World Wide Web Conf.*, New York, NY, USA, May 2019, pp. 1035–1038.

[80] D. Gayo-Avello, "'I wanted to predict elections with Twitter and all I got was this lousy paper'—A balanced survey on election prediction using Twitter data," 2012, *arXiv:1204.6441*.

[81] Mumbling. (Dec. 2017). *Ratioed*. [Online]. Available: https://www.urbandictionary.com/define.php?term=Ratioed

[82] Merriam–Webster. (2017). *Words We're Watching: 'Ratioed'*. [Online]. Available: https://www.merriam-webster.com/words-at-play/words-were-watching-ratio-ratioed-ratioing

[83] (May 2021). *Ratio*. [Online]. Available: https://www.dictionary.com

[84] J. R. Minot, M. V. Arnold, T. Alshaabi, C. M. Danforth, and P. S. Dodds, "Ratioing the president: An exploration of public engagement with Obama and Trump on Twitter," 2020, *arXiv:2006.03526*.

[85] P. A. Longley, M. Adnan, and G. Lansley, "The geotemporal demographics of Twitter usage," *Environ. Planning A, Economy Space*, vol. 47, no. 2, pp. 465–484, Feb. 2015.

[86] Z. Shah, P. Martin, E. Coiera, K. D. Mandl, and A. G. Dunn, "Modeling spatiotemporal factors associated with sentiment on Twitter: Synthesis and suggestions for improving the identification of localized deviations," *J. Med. Internet Res.*, vol. 21, no. 5, May 2019, Art. no. e12881.

[87] A. Ohlheiser, "How K-pop fans became celebrated online vigilantes," MIT Technol. Rev., Cambridge, MA, USA, Tech. Rep., Jun. 2020. [Online]. Available: https://www.technologyreview.com/2020/06/05/1002781/kpop-fans-and-black-lives-matter/

[88] E. Le Merrer, B. Morgan, and G. Trédan, "Setting the record straighter on shadow banning," in *Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM)*, May 2021, pp. 1–10, doi: 10.1109/INFOCOM42981.2021.9488792.

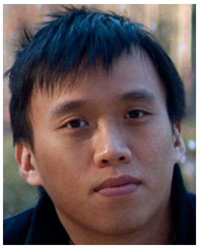
[89] R. T. Garcia. (Apr. 2020). *Harmful Tweets from High Places: Why is Twitter Acting Now?* Salon. [Online]. Available: https://www.salon.com/2020/04/25/harmful-tweets-from-high-places-why-is-twitter-acting-now_partner/

[90] K. Lyons, "Twitter censored tweets critical of India's handling of the pandemic at its government's request," Verge, Apr. 2021. [Online]. Available: https://www.theverge.com/2021/4/24/22400976/twitter-removed-tweets-critical-india-censor-coronavirus

[91] E. Le Merrer, B. Morgan, and G. Trédan, "Setting the record straighter on shadow banning," 2020, *arXiv:2012.05101*.

[92] S. Zannettou, "'I won the election!': An empirical analysis of soft moderation interventions on Twitter," 2021, *arXiv:2101.07183*.

[93] N. P. Suzor, S. M. West, A. Quodling, and J. York, "What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation," *Int. J. Commun.*, vol. 13, pp. 1526–1543, Mar. 2019.



JWEN FAI LOW received the M.Sc. degree in computing and information science from the Masdar Institute of Science and Technology (now part of Khalifa University), United Arab Emirates. He is currently pursuing the Ph.D. degree with the School of Information Studies, McGill University, Canada. His current research interests include data mining, natural language processing, social cybersecurity, and agent-based modeling of information diffusion on social media.

BENJAMIN C. M. FUNG (Senior Member, IEEE) received the Ph.D. degree in computing science from Simon Fraser University, Canada, in 2007. He is the Canada Research Chair of Data Mining for Cybersecurity and a Professor with the School of Information Studies, McGill University, Canada. He is a licensed Professional Engineer of software engineering in Ontario, Canada. He has over 140 refereed publications, with more than 14,000 citations and H-index 57, that span the research forums of data mining, privacy protection, cybersecurity, services computing, and building engineering. He serves as an Associate Editor for IEEE Transactions of Knowledge and Data Engineering (TKDE) and *Sustainable Cities and Society* (SCS) (Elsevier).

FARKHUND IQBAL (Member, IEEE) received the master's and Ph.D. degrees from Concordia University, Canada, in 2005 and 2011, respectively. He is a Professor with the College of Technological Innovation, Zayed University, United Arab Emirates. He is an Adjunct Professor with the School of Information Studies, McGill University, Canada, and an Associate Graduate Faculty Member with the Faculty of Business and IT, Ontario Tech University, Canada. He is the Team Lead of the Cybersecurity and Digital Forensics (CAD) Research Group, Center for Smart Cities and Intelligent Systems, Zayed University. He has 15+ years of teaching and research experience. His research interests include artificial intelligence, machine learning, service robotics, and data analytics techniques for problem-solving in digital security, digital forensics, healthcare, and smart city domains.