Self-disclosure on Twitter during the COVID-19 Pandemic: A Network Perspective

Prasanna Umar, Chandan Akiti, Anna Squicciarini(✉), and Sarah Rajtmajer

College of Information Sciences and Technology, Pennsylvania State University,
University Park, PA 16802, USA
{pxu3,cra5302,acs20,smr48}@psu.edu

Abstract. Amidst social distancing, quarantines, and everyday disruptions caused by the COVID-19 pandemic, users' heightened activity on online social media has provided enhanced opportunities for self-disclosure. We study the incidence and the evolution of self-disclosure temporally as important events unfold throughout the pandemic's timeline. Using a BERT-based supervised learning approach, we label a dataset of over 31 million COVID-19 related tweets for self-disclosure. We map users' self-disclosure patterns, characterize personal revelations, and examine users' disclosures within evolving reply networks. We employ natural language processing models and social network analyses to investigate self-disclosure patterns in users' interaction networks as they seek social connectedness and focused conversations during COVID-19 pandemic. Our analyses show heightened self-disclosure levels in tweets following the World Health Organization's declaration of pandemic worldwide on March 11, 2020. We disentangle network-level patterns of self-disclosure and show how self-disclosure characterizes temporally persistent social connections. We argue that in pursuit of social rewards users intentionally self-disclose and associate with similarly disclosing users. Finally, our work illustrates that in this pursuit users may disclose intimate personal health information such as personal ailments and underlying conditions which pose privacy risks.

Keywords: Self-disclosure · Twitter · Privacy.

1 Introduction

The COVID-19 pandemic has impacted a majority of the world population. As of March 2021, more than 117 million people worldwide have been infected by the coronavirus and more than 2.59 million have died. Much of the world has been living with lockdowns and quarantines since the early months of 2020. Amidst these circumstances, people have resorted to online resources to stay connected in their personal and professional lives. As a result, there has been an unprecedented surge in online activity. Social media usage has increased by 61% as people converge to these online platforms to support their social interactions [25]. Twitter, a popular microblogging site, has seen substantial increase in number of active users during the pandemic [2].

The convergence of people to social media, particularly micro-blogging sites like Twitter, is an evident phenomenon during natural disasters (e.g., earthquakes, hurricanes, floods) and social change events (e.g., black lives matter, occupy wall street) [28, 31]. Amidst the heightened online activity, users can disclose sensitive and private information. In fact, existing literature on social media use during disasters has maintained that a significant portion of user messaging is of a personal nature [27, 42]. Some informational and emotional disclosures are relevant to raising situational awareness during these events and helping in response (e.g., location, life and property loss, mental states) [27]. But, instances of personal disclosures not directly relevant to disaster response have also been observed [27]. Therefore, it is yet unknown how this behavior is different from usual sharing practices. Further, it is unclear how to characterize self-disclosure in health-related crises.

Notably, the COVID-19 pandemic as a health crisis is different than other types of disasters given the scale of the crisis and the global restrictions on movement it has brought for such an extended period of time. Amidst concerns of financial security, health risks and social isolation [35], social media provides an avenue for "collective coping" wherein users seek and receive emotional, informational and instrumental support [28]. Individuals find a sense of community online and feel supported through sharing with others co-experiencing similar problems. Online sharing serves therapeutic functions [18] and enables sense-making in stressful crisis situations [28]. In addition, stressful life events have been shown to mitigate privacy concerns linked to self-disclosure in online social networks [48, 49]. Accordingly, we suggest that users curate their social connections and disclose intentionally to reap social benefits during difficult times.

In this work, we seek to detect levels of self-disclosure in users' public tweets related to the COVID-19 pandemic and characterize these disclosures. We categorize self-disclosure in tweets along several dimensions, namely, information, thoughts, feelings, intimacy and relationships. Leveraging a BERT-based automated labelling scheme trained on human annotations, we assess levels of self-disclosure and its dimensions in more than 31 million tweets. We analyze the labelled data to characterize the phenomenon of self-disclosure during the COVID-19 pandemic. Our work is guided by the following research questions.

– RQ1: What sharing patterns characterize the interaction networks in Twitter and how do these patterns evolve temporally?
– RQ2: Does self-disclosure aid in fostering persistent and focused social interactions?
– RQ3: What content characterizes health-related disclosures among temporally persistent user interactions during the pandemic?

Our findings support the role of self-disclosure in soliciting social connectedness and curating support networks during the pandemic. Our analyses provide several important insights that lead to the following observations. First, we observe heightened self-disclosure levels in tweets following the World Health Organization's March 11 2020 pandemic declaration, signaling a shift in users' sharing patterns in step with heightened awareness and anxiety around the crisis. Second, we find that self-disclosure levels remain consistently high many months into the pandemic. Our analyses of users' reply networks yield novel insights into the temporal evolution of user groups in terms of self-disclosure levels and topical conformity. Specifically, users' interactions within reply networks show more frequent and more intimate self-disclosures temporally. Further, users tend to connect with other users with similar self-disclosure levels, i.e., users show assortative [34] behavior in terms of sharing patterns. Self-disclosures appear to foster more focused and on-topic conversations. Finally, health-related conversations among users include disclosures of personal ailments and health conditions signaling shifts in users' risk perceptions towards sensitive personal health information (PHI) and engagement in such disclosures for potential social rewards.

2 Dataset

The dataset used in our analyses is a subset of a recently collected COVID-specific Twitter repository [14]. The original repository consisted of about 508 million tweet IDs. These tweet IDs corresponded to tweets that were collected using a specific set of keywords (e.g., Coronavirus, CDC, COVID-19, pandemic, SocialDistancing, quarentinelife, etc.) and by following a set of accounts focused on COVID-19 (e.g., CoronaVirusInfo, V2019N, CDCemergency, CDCgov, WHO, etc.). Around June 6th, there was a significant increase in volume of the tweets collected because of changes in collection infrastructure. The transition, however, did not result in any gaps within the timeline of the collected tweets (See [14] for details). Using Python's Twarc package, we re-hydrated the tweets from the tweet IDs. We considered only the original content (English) posted by users i.e, replies and filtered all retweets and quotes. It has been noted that users with high number of followers do not necessarily reciprocate interactions from other users [29]. Highly followed users, by this measure, are not necessarily the most important in the network. Hence, we also removed all tweets from verified accounts. Direct replies to tweets from verified accounts were also removed to exclude the mostly non-reciprocated and one-way interactions within the Twitter network. Similarly, we found majority of user mentions to be targeted towards verified accounts and were not reciprocated. Therefore, we removed user mentions. Our resulting corpus contained just over 31 million tweets, collected from 1/21/2020 till 8/28/2020. We grouped the tweets temporally into three phases. Division into phases was done to test temporal changes in self-disclosing trends that occurred as a result of real-world events related to the pandemic. Specifically, we considered the World Health Organization's declaration of the global pandemic on March 11, 2020 as a "starting" point for the pandemic [41]. Similarly, we selected July 1, 2020 as the beginning of Phase III in our data to reflect the relative easing of strict quarantine and travel restrictions [1]. Accordingly, Phases I (Jan 21 - Mar 11), II (Mar 12 - Jun 30), and III (Jul 1 - Aug 28) comprised of over 4.18 million, 11.83 million, and 15 million tweets respectively.

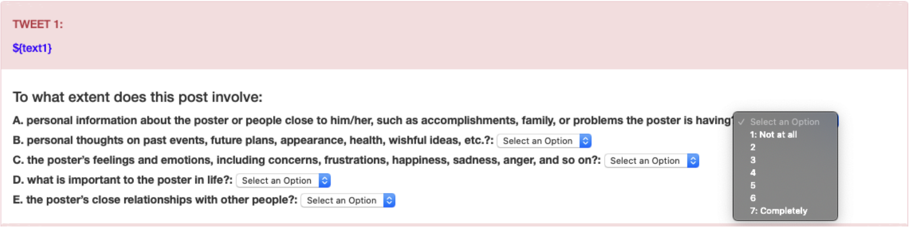
3 Self-disclosure Measurements

3.1 Measurement Scale

We adopted an existing self-disclosure scale [46] to measure level of personal disclosure in tweets[1]. Self-disclosure is operationalized per this measurement scale as a composite value of five items, each measured on an integer scale between 1 (`not at all`) and 7 (`completely`), where 1 represents no disclosure and 7 is the highest level of self-disclosure. Individual items within this framework measure disclosure of: personal information; personal thoughts; personal feelings and emotions; importance/intimacy of the disclosure; and, disclosure of close relationships (See Figure 1 for details).

3.2 Manual Annotations

We labelled a sample of 5000 tweets for self-disclosure using the survey in [46]. The labelling survey was deployed on Amazon Mechanical Turk where each tweet was labelled by three crowd-sourced raters. The labeling task was conducted under the protocol 14947 approved by the Pennsylvania State University's Institutional Review Board (IRB). To ensure quality labels, we provided detailed instructions and examples in the survey. Raters were asked to label each tweet along the five dimensions of self-disclosure considering only the text of the tweet. We, therefore, replaced the weblinks in the tweets with a token `:URL:` and replaced any emoticon with its textual version. We authorized workers only in United States with at least 98% of their past submissions and at least 100 submissions accepted. Further, we discarded responses (about 1% of total submissions) from workers who failed to answer an attention check question within the survey.



TWEET 1:

${text1}

To what extent does this post involve:

A. personal information about the poster or people close to him/her, such as accomplishments, family, or problems the poster is having

B. personal thoughts on past events, future plans, appearance, health, wishful ideas, etc.?: Select an Option

C. the poster's feelings and emotions, including concerns, frustrations, happiness, sadness, anger, and so on?: Select an Option

D. what is important to the poster in life?: Select an Option

E. the poster's close relationships with other people?: Select an Option

Fig. 1. Labelling survey for a tweet showing five questions that represent five dimensions of self-disclosure.

The crowd-sourced workers rated each tweet on an integer scale from 1 (Not at all) to 7 (Completely) for presence of self-disclosure according to each of the five dimensions – namely information, thoughts, feelings, intimacy and relations (See

[1] Authors of [46] reported a reliability of 0.72 (Cronbachś alpha) for the scale

Table 1. Inter-rater agreement for self-disclosure dimensions.

Items	Gwetś AC2	95% CI	Percent Agreement	Benchmark
Information	0.869	0.860 - 0.877	0.922	almost perfect
Thought	0.258	0.240 - 0.276	0.797	fair
Feeling	0.651	0.636 - 0.666	0.859	substantial
Intimacy	0.849	0.842-0.856	0.905	almost perfect
Relation	0.971	0.969-0.974	0.975	almost perfect

Figure 1). For each of these individual ratings, we calculated Gwet's AC2, a chance-corrected agreement statistic [24]. As the individual dimensions of self-disclosure were measured on an ordinal scale, we used the weighted version (ordinal) of Gwet's AC2 statistic and interpreted the magnitude using a bench-marking procedure in [24]. Agreement between raters varied for individual dimensions ranging from fair agreement for `thought` to better agreement for other dimensions (See Table 1). Ratings for each dimension were calculated by averaging the ratings provided by three raters. A final self-disclosure rating was compiled as an average of ratings across the five individual dimensions.

3.3 Label Generation

We generated labels for an unlabelled tweet using labelled examples of tweets in each of the five dimensions of self-disclosure: information, thought, feeling, intimacy and relation. We built separate models for each dimension and aggregated the ratings for all five dimensions to get a self-disclosure rating.

We formulate the labeling process as a regression problem. Formally, we learn a model $h_\theta(x)$ from a set of $(N_u + N_l)$ training samples, where N_u and N_l are the number of unlabeled and labeled examples respectively. The labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ where $y_i \in [1, 7]$ is a small dataset of 5000 samples. We use few-shot learning method [16] to give the model the ability to label unseen samples with only a few labeled known samples. Our learning model is the transformer-based language model called BERT [16]. This learning model has state-of-the-art performance on several standard NLP tasks [44] that closely relate to our regression problem. Thus, BERT is extremely suitable for transferring the learnt knowledge θ to our regression problem.

Domain Fine-Tuning. Following [23], we fine-tuned the pre-trained BERT again on the huge unlabeled data $\mathcal{D}_u$. We use the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives to fine-tune the BERT model. Fine-tuning the language model on target domain data improves performance of NLP tasks and we observe the same in our results in Table 2.

Training Method. In a standard supervised training method, we sample a batch of b samples $\mathcal{B}_r = \{(x_{r_i}, y_{r_i})\}_{i=1}^{b}$ sampled randomly from the N_l labeled

Table 2. RMSE and accuracy (weighted) within ±0.5 and ±1 ranges of true labels.

Label	BERT-base			Fine-tuned			Few-Shot		
	rmse	± 0.5	± 1.0	rmse	± 0.5	± 1.0	rmse	± 0.5	± 1.0
Information	1.689	46.4	63.5	1.213	51.3	68.1	1.027	57.2	74.1
Thoughts	1.827	27.8	45.4	1.653	31.2	47.3	1.406	33.2	54.9
Feelings	1.522	47.8	59.1	1.454	46.9	58.7	1.339	49.6	64.0
Intimacy	0.903	59.8	88.9	0.855	67.2	90.1	0.921	87.6	88.8
Relation	0.639	88.3	94.7	0.632	88.1	94.9	0.586	94.8	95.7

samples. We then pass the training batch $\mathcal{B}$ to the model h_θ to obtain the outputs $\{\hat{y}_{r_i}\}_{i=1}^{b}$. The parameters θ are trained with the loss $\frac{1}{b}\sum_{i=1}^{b}(y_{r_i} - \hat{y}_{r_i})$. As the data-imbalances add huge bias to the model, we re-sample the training samples to balance the samples for each class or ranges.

The few-shot learning method [8] learns to predict labels using a support sample set as knowledge. This learning paradigm works well in low-resource settings. We sample episodes instead of batches, where each episode has a support batch $\mathcal{B}_s = \{(x_{s_i}, y_{s_i})\}_{i=1}^{b}$ and query batch $\mathcal{B}_q = \{(x_{q_i}, y_{q_i})\}_{i=1}^{b}$. Every batch we sample has nearly equal representation from all label classes/ranges. The regression layer $\theta_R \subset \theta$ is removed in this model. Instead, θ_R is inferred from the sentence representations of support set samples. In each episode θ_R is learned with respect to $\mathcal{B}_s$ and then used to predict the labels on $\mathcal{B}_q$. The regression loss is calculated similarly to the supervised learning method and model parameters θ are updated using back-propagation.

We set batch size b to 50, as a lower batch size leads to instability in solving for θ_R. The BERT model outputs sentence representations of dimension 768. We train the model for 3 epochs. Our batch sampling strategy ensures equal number of samples in the six range spans for labels in [1.0-7.0] that are – [1.0-1.5), [1.5, 2.5), [2.5, 3.5), [3.5, 4.5), [4.5, 5.5), [5.5, 7.0].

Evaluation. We use two baseline models to evaluate the performance of our method – namely a standard pre-trained *bert-base* model and a *bert-base* model fine-tuned on the unlabeled dataset. Both baselines are trained with batch size of 32 with a learning rate of $5e-4$ and for 2 epochs. For each sample, we assign an appropriate range/class based on the true label (e.g., [1.5, 2.5) is the range for true label 2.3). Then, the sample prediction is evaluated for this range as a true positive if it falls within margins of ±0.5 (and ±1.0) of the selected range. The results shown in Table 2 indicate that average performance of our method is better than the baseline models.

4 Analysis

In this section, we describe our methodology to understand patterns of self-disclosure during the Covid-19 pandemic, and present our findings. We construct both directed and undirected reply-based graphs wherein users are represented as vertices and pairwise reply interactions between users are represented as edges. We posit that for our study of predominantly conversation-oriented behaviors such as self-disclosure a suitable representation of the system is a network that captures reply-based interactions between users. While studies often characterize Twitter as a static network [6, 29], it has been acknowledged that such networks can be misleading [22, 26]. The follower/following relationships are mostly not reciprocated and follower/following-based networks do not give actual representations of users' active reciprocated interactions [26].

4.1 Self-disclosure Assortativity in Twitter Reply Networks

In order to understand the self-disclosure patterns that characterize the Twitter interaction network (RQ1), we examine if users' sharing patterns are similar to their social connections. That is, we peruse the assorativity of users' interaction networks in terms of self-disclosure patterns.

Reciprocal-reply Network. We create reciprocal-reply networks [11] to explore the assortative mixing of users according to their patterns of self-disclosure. Particularly, we define an undirected graph $G(V, E)$ with a set of vertices, V and a set of edges pairwise amongst them, E. For users $v_j \in V$ and $v_k \in V$, an edge e_{jk} represents a reciprocal reply relationship between them, i.e., existence of replies by both users to each other. Each edge is assigned a weight w_{jk} calculated as the sum of number of interactions (replies) between two users v_j and v_k. Three reciprocal-reply networks were created to represent each temporal Phase in the dataset. Here, mean self-disclosure characterizes each node as an attribute and it is calculated as the average of self-disclosure levels across all tweets posted by the user (node) within the particular phase. Essentially, two nodes v_j and v_k connected by an undirected edge e_i had attributes j_i and k_i. We calculate assortativity based on average self-disclosure using a weighted version of the continuous assortativity coefficient in [21]. Specifically, assortativity coefficient is defined using the equation 1 where W is the sum of all edge weights. The values for assortativity coefficient range from -1 to 1. Positive values of this coefficient means similarities among connected nodes and dissimilarities results in negative values.

$$r_c^w = \frac{\sum_i w_i j_i k_i - W^{-1} \sum_i w_i j_i \sum_{i'} w_{i'} k_{i'}}{\sqrt{[\sum_i w_i j_i^2 - W^{-1} \sum_i w_i j_i^2][\sum_i w_i k_i^2 - W^{-1} \sum_i w_i k_i^2]}} \tag{1}$$

Results. We found evidences of assortativity for mean self-disclosure among users (See Table 3). For the first Phase, the network had negligible but positive assortativity coefficient (0.003) which increased for Phase II (0.239) and Phase III (0.218). While there are no formal guidelines for interpreting assortative coefficient, we follow the most recent study that re-purposes correlation coefficient ranges to classify networks into levels of assortativity [30]. Accordingly, the first Phase

Table 3. Assortative coefficient for reciprocal reply networks.

Phase	#Nodes	#Edges	SD	Information	Thought	Feeling	Intimacy	Relation
Phase I	9474	5479	0.003	0.165	-0.001	-0.002	-0.014	-0.005
Phase II	8201	4334	0.239	0.245	0.313	0.259	0.147	0.156
Phase III	25372	14216	0.218	0.263	0.280	0.248	0.116	0.148

network is interpreted as neutral (i.e., neither assortative nor disassortative) and the networks in subsequent phases are considered to be weakly assortative. Along the dimensions of self-disclosure, similar increasing patterns were observed. However, the reciprocal-reply networks in all three phases are neutral for feeling and intimacy dimensions of self-disclosure.

4.2 Persistent Groups and Self-disclosure

In order to understand how the occurrences of self-disclosure characterize and aid in the persistent and focused social interactions (RQ2), we examine the temporally persistent social groups. For these groups of users, we peruse temporal evolution of self-disclosure and the relationships with topical conformity.

Directed Reply Networks. We use directed sub-networks to extract self-disclosure patterns and the relationships between self-disclosure and the topical conformity (divergence) within the social connections that temporally persistent groups of users maintain. Specifically, we define a graph $G(V, E)$ on vertex set V and edge set E. For a user $v_i \in V$, edge e_{ij} represents a reply by user v_i to user v_j. Each directed edge e_{ij} is assigned a weight w_{ij} representing the number of these replies. Three such graphs were created, one for each temporal Phase in the data and each graph consisted of all reply interactions between all interacting users within that Phase. We then detect communities in the graphs associated with each Phase using a directed Louvain community detection algorithm optimized for directed modularity [17]. Higher values closer to 1 for modularity indicates stronger community structure and as such, the directed modularity scores for three phases in our study were 0.93, 0.98 and 0.94 respectively. We identify persistent groups of (at least 2) users which interact within same communities across three phases. That is, a persistent group represents a set of users which is a part of a larger community in Phase I and persisted as a group within common communities across subsequent phases, although the group as a whole can be a part of different communities in subsequent phases as communities evolve. In total, we pulled out 549 persistent groups totalling 13469 users. Figure 2 shows an example of a persistent group across three phases of the pandemic timeline. For these persistent groups, we examined self-disclosing behaviors across three phases. Additionally, we disentangled the relationship between self-disclosure and the tightness of conversational content posted by the set of users in the persistent group as measured by topical divergence.

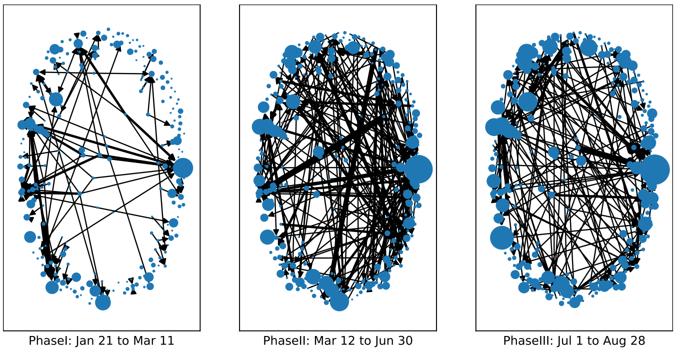


Fig. 2. Directed reply networks for an exemplarly persistent group across phases. Node size is proportional to activity of the user and edge width is proportional to number of replies. Average self-disclosures per phase are 1.18, 1.30, and 1.27 respectively.

Topical Divergence. We perform topical modeling (using Latent Dirichlet Allocation (LDA) [10]) of the tweets from all users in all the persistent groups in order to understand if there is a relationship between self-disclosure in temporally persistent groups and topical conformity in their conversations. We removed hashtags, user mentions, weblinks, and emoticons. We also removed words that appeared in over 90% of the tweets and those that appeared in less than 20 tweets. Multiple topic models were created for a corpus of pre-processed tweets with number of topics varying from 1 to 20. We used coherence score [38] as a measure of quality and interpretability of the topic models. Our extracted best topic model included 17 topics and a coherence score of 0.50 (See Table 4). According to this 17-topic model, we assigned a latent topic distribution vector for each tweet representing probabilities corresponding to each of the topics.

For persistent groups across phases, we measure conformity or lack thereof in conversational content across group members by means of topical divergence. Specifically, we computed the Jensen-Shannon divergence (JSD) for each persistent group across the three phases using the following formulation [37]:

$$JS(g^s) = H(\beta_g^s) - \frac{\sum_{t \epsilon T_g^s} H(\beta_t)}{|T_g^s|} \tag{2}$$

where $\beta_g^s(i) = \frac{\sum_{t \epsilon T_g^s} \beta_t(i)}{|T_g^s|}, \forall i = 1, ..., n$ for n the number of topics is the mean topic distribution of group g at Phase s over all its users' tweets (T_g^s). Here, β_t is the latent topic distribution of tweet t and H is the Shannon-entropy function (logarithmic base 2). The divergence score ranges from 0 to 1 with 1 being totally conforming conversation.

Results. Here, we present our findings on temporal evolution of self-disclosure for persistent groups. Also, we report on the relationship between self-disclosure and topical conformity in the content by the users within these groups.

Table 4. Top keywords for topics generated using LDA with coherence score of 0.50.

1	home, reopen, close, open, stay	10	health, outbreak, warn, spread, travel
2	mask, wear, face, people, social_distancing	11	news, live, update, late, australia
3	cdc, government, control, datum, expert	12	school, child, family, student, kid
4	test, positive, testing, result, symptom	13	like, look, good, time, day
5	people, know, bad, think, die	14	trump, president, response, white_house, election
6	death, case, number, rate, toll	15	vaccine, study, scientist, new, drug
7	crisis, million, business, pay, government	16	fight, india, th, june, july
8	case, new, report, death, total	17	market, economy, fear, hit, amid
9	patient, hospital, die, doctor, care		

Disclosure Patterns. For 549 persistent user groups across three phases, there was a significant difference in self-disclosure rating ($\chi^2(2) = 469.93, p < 0.001, W = 0.428$). Post-hoc analysis with Friedman-Conover tests and Holm-Bonferroni correction revealed significant differences across all phase pairs ($p < 0.001$). Mean values of self-disclosure for three phases were 1.19, 1.38 and 1.35 respectively. It is in line with the overall trend in self-disclosing behavior across all tweets in the dataset through three phases (See Figure 3).

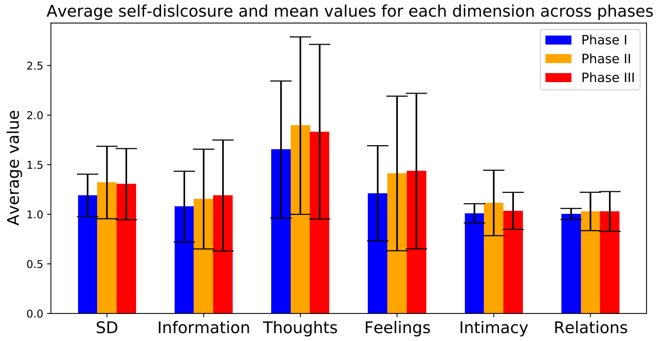


Fig. 3. Average values of SD and its dimensions across phases.

Topical Divergence. We find significant negative correlation between topical divergence and self-disclosure in Phase II and Phase III . Correlations for consecutive phases were -0.11 ($p < 0.05$), -0.55 ($p < 0.001$) and -0.49 ($p < 0.001$) respectively. These findings show that as self-disclosure increases, conversations are more focused and on-topic.

4.3 Characterizing Sensitive Disclosures in Temporally Persistent Social Connections

As users maintained social connections through the pandemic with parallel increases in sharing behavior, we delve further into the content of the disclosures. COVID-19 being a health related crises, we seek to answer RQ3 and characterize the sensitive health disclosures within the persistent social connections.

Sensitive Health-Related Disclosure. We analysed tweets for specific types of health-related disclosure, namely, disclosure of symptoms and diseases. We looked for these specific revelations within the tweets of persistent group members that were classified as having some level (> 1) of informational self-disclosure. To extract these fine-grained utterances, we created a supervised learning model to classify health related tweets. We used an existing manually annotated Twitter dataset [36] with labels specifying health-related content for training purposes. The dataset contained 5128 tweet IDs corresponding to tweets that were labeled according one of five categories: sick, health, unrelated, not English, and ambigous. Excluding the tweets that could not be retrieved, non-English and ambiguous tweets, we compiled 2419 tweets. We binarized the dataset into 987 health-related tweets (sick, health) and 1432 non-health related tweets. We use our first baseline BERT based model to train on this dataset and we infer a binary health-related vs non-health related label for all the tweets from all persistent groups. We use the same hyperparameters used in our labeling baseline. The model trained with 5-fold cross-validation yielded average (validation) precision, recall and F1-score of 78.8%, 84.4% , and 81.4% respectively.

We used a pre-trained model [40] to detect disclosures of symptoms and diseases in health-related tweets that were tagged as containing at least some (> 1) levels of information disclosure. Authors of [40] trained and evaluated this model on a Twitter dataset and reported 72% F1-score for detection of medical entities. Using the trained model, we detected the sensitive disclosures of symptoms and diseases within the informational disclosures in the health-related tweets by the persistent groups.

Results. Topics of conversation within persistent groups highlighted health-related discussions. About 29% of all tweets within persistent groups were tagged as health-related and 83% of the persistent groups had at least some health-related tweets. Notably, 99.7% of these health related tweets belonged to the seven topics that featured health-related keywords (See topics 3-6, 8-9, 15 in Table 4). Zooming in on health-related tweets that had at least some informational disclosure (rating > 1), we detected disclosures of personal ailments and symptoms (see Table 5).

5 Discussion

Our analyses showed increased levels of self-disclosure in COVID-19 related tweets after March 11, 2020 when the WHO declared the outbreak a global

Table 5. Examples of disclosures of personal ailments.

’Damn. 1. I have a cold. 2. I have not been to China. 3. I have travelled in the last week. Once to London. How worried should I be?’	’cold’
’Wondering if the sore throat I developed this afternoon is the coronavirus. I guess we shall soon see.’	’sore throat’, ’coronavirus’
’I though i will be fine at ome spend the dya sleeping yesterday and now woke up with a head ache again and diff breathing before going to see a cardiologist i need to pass a test to eliminate f* * *ing covid do not want to go too tired’	’head ache’, ’breathing’
’I have very little positivity to share I am afraid today. My Son is visiting & we are going out with Dogs. I am concerned for him, he has the same Kidney Condition as I, he inherited before I knew I had from my Dad. My Girls &Grandchildren as is in’	’Kidney Condition’

pandemic (also observed in recent work [41]). This increase, registered both in terms of quantity and intimacy of self disclosure, coincided with acute temporal events in pandemic timeline and suggests that self-disclosure has served an important role ameliorating social and emotional challenges linked with the crisis. Users have turned to online communities for support [13]. Recent work studying potential changes in individual perceptions of self-disclosure and privacy during the pandemic [33] supports this view.

Reciprocal-reply networks reveal assortative mixing of users based on self-disclosure behavior after March 11. Such self-organized mixing patterns in online social networks as a result of acute disaster has been observed in recent work [19]. Authors of [19] showed that (degree) assortative mixing patterns vary with evolution of disaster as critical events unfold and emergent social cohesion is intentional in pursuit of specific needs. We suggest that stresses of the pandemic may have likewise enabled selective mixing of users in terms of self-disclosure, following work on the role of self-disclosure in maintaining relationships and psychological coping [3]. Our results provide initial evidence of users’ curation of social connections and strategic self-disclosure in pursuit of social rewards.

We have also shown that self-disclosure by users within temporally persistent social groups supports focused, on-topic conversations, highlighting the role of self-disclosure in maintaining stable support structure. Further studies could delve deeper into these effects in emergent support-oriented communities, particularly in crisis.

Amongst users within persistent social groups, we found disclosure of sensitive personal health information (PHI) such as physical ailments, symptoms and underlying health conditions. While observed in dedicated online health communities (OHC) [47], such sensitive voluntary disclosures in Twitter during crises is relatively under-studied. Studies on OHC show that pursuit of informational and emotional support motivates PHI disclosures [50]. We speculate that users in our dataset similarly disclosed sensitive PHI to garner support from their Twitter community. As noted by [33], the pandemic may have changed privacy perceptions towards sensitive PHI. Additional work in this area could shed light on the motivations for and differences in PHI disclosures in user engagements during crises vs normal times.

6 Related Work

Since 2020, a body of literature has emerged studying activity in Twitter to understand user sentiment [39], explore prevalence and prevention of misinformation [45], and analyze hate speech [20] during the pandemic. Often, these studies perform raw tweets collection, conduct content analyses, and build models to answer specific research questions related to trends in online social behavior. As a result, over the past year several Twitter datasets [7, 15] and computational models [32, 51] have been released. Yet, studies to date have not focused on analyses of the extant network in which these trends occur. Further, we are not aware of any study that looks into network effects on self-disclosure during the pandemic. We attempt to fill these gaps.

Outside the domain of crisis informatics, self-disclosure has been studied as an intentional and influenced behavior which has both intrinsic and extrinsic rewards [3]. Intrinsically, it has therapeutic benefits that can help in psychological well-being [43] and extrinsically, it plays a role in building relationships, social connectedness, and maintaining relationships [4]. Increasingly, studies have perused self-disclosure in social networking sites where users look to interact with others for both intrinsic and extrinsic benefits. However, we find differing approaches for operationalization and measurement of self-disclosure throughout the literature [3]. Of interest for observational studies, [9] proposed a 3-item scale (levels of information, thoughts and feelings) to measure self-disclosure in online posts. However, we follow a more recent work [46] that modified this scale to include the intimacy of disclosure.

Similar to [46], most studies create automated models to scale self-disclosure labels in small manually annotated data to larger samples [5, 12]. Such models employ highly curated dictionaries and extensive feature engineering which limit the inference process and performance on unseen data. Here, we use transfer-learning techniques on NLP models for labeling of our self-disclosure text.

7 Conclusion

Our study sheds light on the increase in users' self-disclosure during the pandemic and the role of self-disclosure in persistent and transient online groups. We have suggested that users share personal information in their online communities to garner social support. Reinforcing this argument, our results showed that as users maintained social connections temporally, self-disclosure increased as did topical conformity within conversations. Disclosures of users within persistent groups revealed sensitive personal health information. As such, our study points toward shifts in users' privacy perceptions in the wake of the COVID-19 pandemic.

As our findings are empirical in nature, a limitation of our work relates to the data we rely on. Although the dataset captures tweets in the important timeline of the pandemic, it is a sample of COVID-19 related conversations on Twitter. Hence, the results of this study need to be interpreted accounting for the effects of missing data in the sample.

Acknowledgements

Work from all the authors was supported in part by the National Science Foundation under Grant 2027757.

References

1. Coronavirus: How lockdown is being lifted across europe Accessed: 2021-03-08
2. Twitter sees record number of users during pandemic, but advertising sales slow Accessed: 2021-03-08
3. Abramova, O., Wagner, A., Krasnova, H., Buxmann, P.: Understanding self-disclosure on social networking sites - A literature review. In: AMCIS 2017 Proceedings. pp. 1–10. No. August (2017)
4. Aharony, N.: Relationships among attachment theory, social capital perspective, personality characteristics, and facebook self-disclosure. Aslib Journal of Information Management (2016)
5. Bak, J., Lin, C.Y., Oh, A.: Self-disclosure topic model for classifying and analyzing twitter conversations. In: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1986–1996. Association for Computational Linguistics, Doha, Qatar (Oct 2014)
6. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 65–74 (2011)
7. Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Chowell, G.: A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. arXiv preprint arXiv:2004.03688 (2020)
8. Bao, Y., Wu, M., Chang, S., Barzilay, R.: Few-shot text classification with distributional signatures (2020)
9. Barak, A., Gluck-Ofri, O.: Degree and reciprocity of self-disclosure in online forums. CyberPsychology & Behavior **10**(3), 407–417 (2007)
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
11. Bliss, C.A., Kloumann, I.M., Harris, K.D., Danforth, C.M., Dodds, P.S.: Twitter reciprocal reply networks exhibit assortativity with respect to happiness. Journal of Computational Science **3**(5), 388–397 (2012)
12. Caliskan Islam, A., Walsh, J., Greenstadt, R.: Privacy detective: Detecting private information and collective privacy behavior in a large social network. In: 13th Workshop on Privacy in the Electronic Society. pp. 35–46. ACM (2014)
13. Chakraborty, T., Kumar, A., Upadhyay, P., Dwivedi, Y.K.: Link between social distancing, cognitive dissonance, and social networking site usage intensity: a country-level study during the covid-19 outbreak. Internet Research (2020)
14. Chen, E., Lerman, K., Ferrara, E.: COVID-19: The First Public Coronavirus Twitter Dataset. arXiv e-prints arXiv:2003.07372 (Mar 2020)

15. Chen, E., Lerman, K., Ferrara, E.: Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR Public Health and Surveillance **6**(2), e19273 (2020)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
17. Dugué, N., Perez, A.: Directed Louvain: maximizing modularity in directed networks. Ph.D. thesis, Université d'Orléans (2015)
18. Ernala, S.K., Rizvi, A.F., Birnbaum, M.L., Kane, J.M., De Choudhury, M.: Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. Proceedings of the ACM on Human-Computer Interaction **1**(CSCW), 1–27 (2017)
19. Fan, C., Jiang, Y., Mostafavi, A.: Emergent social cohesion for coping with community disruptions in disasters. Journal of the Royal Society Interface **17**(164), 20190778 (2020)
20. Fan, L., Yu, H., Yin, Z.: Stigmatization in social media: Documenting and analyzing hate speech for covid-19 on twitter. Proceedings of the Association for Information Science and Technology **57**(1), e313 (2020)
21. Farine, D.: Measuring phenotypic assortment in animal social networks: weighted associations are more robust than binary edges. Animal Behaviour **89**, 141–153 (2014)
22. Gonçalves, B., Perra, N., Vespignani, A.: Modeling users' activity on twitter networks: Validation of dunbar's number. PloS one **6**(8), e22656 (2011)
23. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks (2020)
24. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
25. Holmes, R.: Is covid-19 social media's levelling up moment? Forbes **24** (2020)
26. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. arXiv preprint arXiv:0812.1045 (2008)
27. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Extracting information nuggets from disaster-related messages in social media. In: Iscram (2013)
28. Jurgens, M., Helsloot, I.: The effect of social media on the dynamics of (self) resilience during disasters: A literature review. Journal of Contingencies and Crisis Management **26**(1), 79–88 (2018)
29. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. pp. 591–600 (2010)
30. Meghanathan, N.: Assortativity analysis of real-world network graphs based on centrality metrics. Computer and Information Science **9**(3), 7–25 (2016)
31. Miyabe, M., Miura, A., Aramaki, E.: Use trend analysis of twitter after the great east japan earthquake. In: ACM 2012 conference on Computer Supported Cooperative Work Companion. pp. 175–178 (2012)
32. Müller, M., Salathé, M., Kummervold, P.E.: Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503 (2020)
33. Nabity-Grover, T., Cheung, C.M., Thatcher, J.B.: Inside out and outside in: How the covid-19 pandemic affects self-disclosure on social media. International Journal of Information Management **55**, 102188 (2020)
34. Noldus, R., Van Mieghem, P.: Assortativity in complex networks. Journal of Complex Networks **3**(4), 507–542 (2015)

35. Ognyanova, K., Perlis, R.H., Baum, M.A., Lazer, D., Druckman, J., Santillana, M., Volpe, J.D.: The state of the nation: A 50-state covid-19 survey report #4 (2020)
36. Paul, M., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: International AAAI Conference on Web and Social Media. vol. 5 (2011)
37. Purohit, H., Ruan, Y., Fuhry, D., Parthasarathy, S., Sheth, A.: On understanding the divergence of online social group discussion. In: International AAAI Conference on Web and Social Media. vol. 8 (2014)
38. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Eighth ACM international conference on Web search and data mining. pp. 399–408. ACM (2015)
39. Sanders, A.C., White, R.C., Severson, L.S., Ma, R., McQueen, R., Paulo, H.C.A., Zhang, Y., Erickson, J.S., Bennett, K.P.: Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of covid-19 twitter discourse. medRxiv pp. 2020–08 (2021)
40. Scepanovic, S., Martin-Lopez, E., Quercia, D., Baykaner, K.: Extracting medical entities from social media. In: ACM Conference on Health, Inference, and Learning. pp. 170–181 (2020)
41. Squicciarini, A., Raitmaier, S., Umar, P., Blose, T.: A tipping point? heightened self-disclosure during the coronavirus pandemic. In: IEEE Second International Conference on Cognitive Machine Intelligence (CogMI). pp. 141–146. IEEE (2020)
42. Takahashi, B., Tandoc Jr, E.C., Carmichael, C.: Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. Computers in human behavior **50**, 392–398 (2015)
43. Tamir, D.I., Mitchell, J.P.: Disclosing information about the self is intrinsically rewarding. Proc. of the National Academy of Sciences **109**(21), 8038–8043 (2012)
44. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2019)
45. Wang, Y., Gao, S., Gao, W.: Can predominant credible information suppress misinformation in crises? empirical studies of tweets related to prevention measures during covid-19. arXiv preprint arXiv:2102.00976 (2021)
46. Wang, Y.C., Burke, M., Kraut, R.: Modeling self-disclosure in social networking sites. In: 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. pp. 74–85. CSCW '16, ACM (2016)
47. Yuchao, W., Ying, Z., Liao, Z.: Health privacy information self-disclosure in online health community. Frontiers in Public Health **8**, 1023 (2020)
48. Zhang, R.: The stress-buffering effect of self-disclosure on facebook: An examination of stressful life events, social support, and mental health among college students. Computers in Human Behavior **75**, 527–537 (2017)
49. Zhang, R., Fu, J.S.: Privacy management and self-disclosure on social network sites: The moderating effects of stress and gender. Journal of Computer-Mediated Communication **25**(3), 236–251 (2020)
50. Zhang, X., Liu, S., Chen, X., Wang, L., Gao, B., Zhu, Q.: Health information privacy concerns, antecedents, and information disclosure intention in online health communities. Information & Management **55**(4), 482–493 (2018)
51. Zong, S., Baheti, A., Xu, W., Ritter, A.: Extracting covid-19 events from twitter. arXiv preprint arXiv:2006.02567 (2020)