The Contribution of Verified Accounts to Self-Disclosure in COVID-Related Twitter Conversations

Tingting Du, Prasanna Umar, Sarah Rajtmajer, Anna Squicciarini

College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, USA
{tvd5455, pxu3, smr48, acs20}@psu.org

Abstract

On Twitter, so-called verified accounts represent celebrities and organizations of public interest, selected by Twitter based on criteria for both activity and notability. Our work seeks to understand the involvement and influence of these accounts in patterns of self-disclosure, namely, voluntary sharing of personal information. In a study of 3 million COVID-19 related tweets, we present a comparison of self-disclosure in verified vs ordinary users. We discuss evidence of peer effects on self-disclosing behaviors and analyze topics of conversation associated with these practices.

Introduction

As the COVID-19 pandemic has ensued, social network activities have exploded, with users resorting to online platforms to support nearly every aspect of their life. On Twitter, like in other social networks (SNs), a small subset of accounts have outsized influence on the broader community. So-called verified accounts represent celebrities and organizations of public interest, selected by Twitter based on criteria for both activity and notability. A recent study of celebrities' online activities during the pandemic examined how celebrities talked about COVID-19 in social media posts and how they used their platforms to motivate followers to respond to the virus (Lookadoo et al. 2021). This small study (only 20 influencers were studied) found that at the beginning of the pandemic, celebrities modeled guidance from the Centers for Disease Control and Prevention and used their platforms to normalize recommended health behaviors. In the same way that verified accounts influence conversation, opinion and trends, our work seeks to understand the involvement and influence of these accounts in patterns of self-disclosure (here forward, SD), namely, voluntary sharing of personal information.

In this paper, we study nearly 3 million COVID-19 related Tweets, and present a large scale longitudinal comparison of SD in verified vs. unverified users. We discuss evidence of peer effects on self-disclosing behaviors and analyze topics of conversations associated with these practices. We address the following research questions:

RQ1: How do verified accounts use SD in COVID-related tweets? Studies of SD have generally disregarded public figures and organizational accounts, as concerns about privacy for these users are less clear. We hypothesize, though, that these users play a role in shaping discourse that supports SD and poses privacy concerns, as their posts can be personal while their account pages serve as a town square.

RQ2: Do verified accounts trigger responses or conversations that include incidence of SD? What type of SD is observed in these conversations? Which topics have been most likely to garner SD in verified tweets?

RQ3: How has SD evolved over time during the pandemic for both verified and unverified accounts? From a topical standpoint, are there topics which consistently manifest personal or potentially controversial conversations?

Our results indicate that verified accounts, as anticipated, play a role in SD broadly and in thoughts-related SD specifically. However, their influence in triggering SD during conversations is minimal, and only evident for selected dimensions of SD. During the pandemic, topics rich of personal observations related to conversations centered on health guidelines, e.g., social distancing, masking.

Dataset

The dataset we use here is selected from a repository of COVID-specific tweet IDs (Chen, Lerman, and Ferrara 2020). The 508 million tweet IDs were collected based on relevant keywords and by following accounts related to COVID-19.[1] Using the Twarc package for Python, we re-hydrated tweet text and metadata from tweet IDs, collecting only original English-language content, i.e., filtering retweets, quotes and non-English language text. Our filtered raw dataset contains 49,035,362 COVID-related tweets from January 21 through August 28, 2020. As our work looks at discourse, we extract conversations from the raw dataset, where a conversation is determined as a sequence of at least two tweets related through reply-based interactions. Our final dataset consists of 2,644,357 tweets distributed across 674,739 conversation threads. We segment our analyses into three temporal phases: (a) January 21 to March 10; (b) March 11 to June 30, the acute early phase of the pandemic; and (c) July 1 to August 28. Table 1 provides statistics of the

[1] https://github.com/echen102/COVID-19-tweetIDs/releases

Phase	Tweets	Verified Users	Unverifed Users	Convers.	Thread length	Verif-led convers.	Unverif-led convers.	Human-led convers.	Org-led convers.
I	502,624	10,321	255,702	153,862	3.3 ± 14.2	54,411	99,010	18,339	36,072
II	657,123	17,957	385,893	192,111	3.4 ± 17.8	78,914	112,634	28,819	50,095
III	1,484,610	26,869	736,940	413,258	3.6 ± 24.0	156,746	255,103	55,639	101,107

Table 1: Dataset Statistics

dataset. Each tweet is labeled by Twitter as ”verified” or ”unverified” based on the status of the corresponding account. Verified users can be organizations or humans. We identify whether a verified account is an organization or human using a deep neural architecture-based demographic inference tool (Wang et al. 2019). In particular, we leverage the text-based M3 model that provides the probability of an account being an organization or human based on the user's screenname, username and account description. We find that 39.5% of all verified user accounts were organizations.

Methods

Labeling self-disclosure. We manually labelled 5000 tweets for self-disclosure based on (Wang, Burke, and Kraut 2016) with Amazon Mechanical Turk under approved IRB protocol [*suppressed*]. Each tweet was rated for the presence of SD in 5 categories (“Information”, “Thoughts”, “Feelings”, “Intimacy”, and “Relations”) by three workers on an integer scale from 1 (Not at all) to 7 (Completely).

We use a RoBERTa-based approach to label the unlabeled tweets according to these 5 categories. We fine-tune the pretrained RoBERTa model on our labeled dataset. Specifically, we append a dropout layer and a linear layer to the architecture and use binary cross entropy for our loss function. We trained the model for 4 epochs with a batch-size of 16 and dropout rate 0.1. We used Adam optimizer with learning rate 1e-5, and a linear learning rate warmup on 6% of the training data. To overcome the imbalanced positive and negative samples, we also apply a weight-based sampling so that both classes have equal probability. For the analyses described in this paper, we binarized these labels to represent the simple presence or absence of self-disclosure.

Topic modeling. We perform topic modeling using BERTopic, a hybrid algorithm that combines state-of-the-art language models and clustering algorithms for topic modeling. Before applying the BERTopic algorithm, we first preprocess tweets by removing URLs, emojis, and mentions. In addition, since the dataset we used collected tweets based on designated keywords, we remove these keywords.The BERTopic algorithm proceeds in four stages. First, we extract twitter conversation embeddings using sentence-transformers (Reimers and Gurevych 2019), a state-of-the-art language model for text similarity.

Second, the BERTopic algorithm uses UMAP (McInnes, Healy, and Melville 2020) for non-linear dimensionality reduction as the conversation embeddings from sentence-transformers are 384-dimensional vectors. Third, BERTopic uses HDBSCAN (Campello, Moulavi, and Sander 2013) to find clusters in the embedding space. Hyperparameters of the model were determined using the optuna hyperparameter optimization framework. Number of nearest neighbors, number of components and minimimum cluster size were set to 90, 16 and 120, respectively. We then perform topic reduction, combining least frequent topics with their most similar topics based on distances between topic embeddings. We select the most frequent 50 topics in each phase for analysis.

We look for overlapping topics over the three phases to detect topic continuity. To do so, we create vector representations of topics using embeddings of each topic's top 20 representative words. Then, we take the weighted average of embeddings in a topic by their c-TF-IDF value. This gives greater emphasis to words that better represent each topic. If the cosine similarity between topics in subsequent phases is greater or equal to 0.8, we treat them as the same topic. We identify 15 overlapping topics, 3 of which see significant growth over the three phases, as shown in Figure 1.

Conversation modeling. We thread together direct replies in our dataset to recreate conversations. The tweet that initiates a conversation is considered the parent. Conversations initiated by verified users are further divided into those initiated by organizational vs. human accounts. The number of conversations of each type, by phase, and the mean/median length of conversations is provided in Table 1. In following analyses, we consider 100,000 randomly sampled conversations initiated by unverified users and 100,000 initiated by verified users. Time of a conversation is identified as time of the parent tweet. Rate of SD is computed as the number of tweets containing SD divided by total number of tweets. For topic modeling of conversations, we consider whether a specific topic, e.g., wearing masks, appears in a conversation. That is, if at least one tweet in a conversation is identified within the “wearing masks” topic, the conversation is considered to be about wearing masks. Accordingly some conversations may be multi-topic. We provide insights on SD/non-SD rates for selected topics.

Findings

RQ1: How do verified (vs. unverified) accounts use SD?
We analyze the rate of SD for verified and unverified users over the three phases. We report our findings in Table 3.

Phase	Information	Thoughts	Feelings
I	χ^2=117.52, $p < 0.01$, V=0.015	χ^2=25023.48, $p < 0.01$, V=0.22	χ^2=7921.87, $p < 0.01$, V=0.13
II	χ^2=836.09, $p < 0.01$, V=0.04	χ^2=55688.73, $p < 0.01$, V=0.29	χ^2=18965.46, $p < 0.01$, V=0.17
III	χ^2=2553.37, $p < 0.01$, V=0.04	χ^2=112907.97, $p < 0.01$, V=0.28	χ^2=41451.98, $p < 0.01$, V=0.17

Table 2: Chi-squared tests for relationship between verified nature of account and self-disclosure categories in tweets.

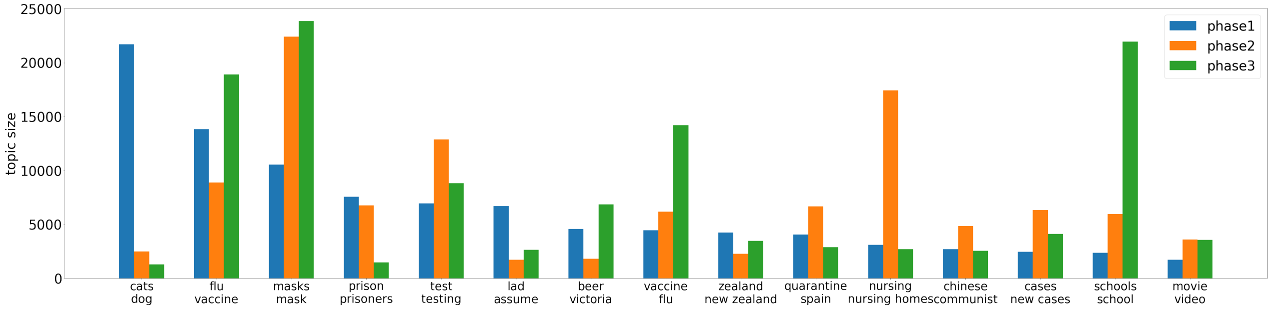


Figure 1: Topical evolution. Some topics saw a multi-fold increase over time, particularly masking and back-to-school.

		I	T	F	I	R
Ph. 1	*ver.*	3.7	32.6	10.0	9.5	0.9
	human	7.8	52.0	23.8	15.1	1.4
	org	1.5	22.1	3.9	6.7	0.1
	unver.	4.7	66.4	28.1	9.5	0.6
Ph. 2	*ver.*	6.0	44.0	14.5	5.1	1.4
	human	11.7	63.6	29.6	8.6	2.5
	org	2.7	32.3	5.7	3.0	0.1
	unver.	9.0	81.2	39.0	8.3	2.3
Ph. 3	*ver.*	6.7	43.5	15.1	4.9	1.8
	human	13.4	63.6	31.3	8.9	2.9
	org	3.0	32.3	6.0	2.8	1.1
	unver.	10.7	80.7	40.9	4.8	2.4

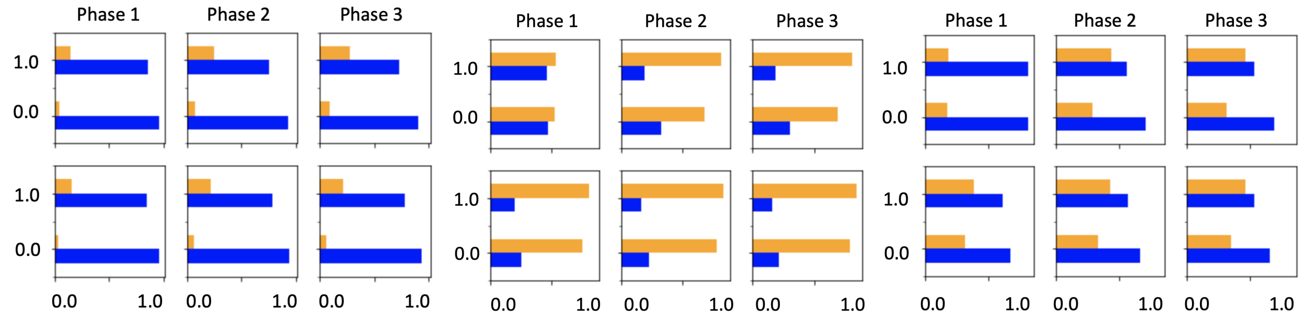
Table 3: SD percentages over time. (I) Information; (T) Thoughts; (F) Feelings; (I) Intimacy; (R) Relations

SD of thoughts is most frequent amongst the 5 categories of disclosure, for all users. Verified human users disclose at a higher rate than unverified users across three of the four other categories – information, intimacy and relations. We perform a t-test to compare the distributions of rate of disclosure between verified human and unverified users. We find that observed differences are statistically significant ($p << 0$) for information, thoughts, feelings and intimacy categories. Differences in SD for the relations dimension are non-significant. Our findings show that even public figures resort to SD to discuss direct experiences with the pandemic. Many public figures have been openly involved in COVID-related conversations, regardless of their profession and reasons for popularity. Anecdotally, celebrities and political figures have been among the first drawing public attention to their COVID-19 diagnoses, updating about health status and normalizing the impacts of restrictions imposed by the pandemic (Lookadoo et al. 2021).

RQ2: Do verified accounts trigger responses or conversations that include incidence of SD? What type of SD is observed in these conversations? Which topics have been most likely to garner SD in verified tweets?

We examined the relationship between SD in a parent tweet and SD in a child tweet during a conversation, across self-disclosure dimensions and conversations of differing length. That is, we are interested in *peer effects* of self-disclosure at two levels. First, we test for peer effects in directs replies (dyadic conversation) to a SD tweet. Then, we examine the proportion of SD in conversations initiated by a SD tweet. We stratify this relationship based on the verified or unverified nature of the parent tweet and further based on the type of verified account.

- SD (information) parent tweets were more likely to be replied to with self-disclosure in child tweets than non-SD tweets, for parent tweets from both verified and unverified users. These relationships were consistently significant across phases. This is true for dyadic conversations (See Figure 2) and also longer conversations. We run a Cramer's test to measure peer effects on informational SD. We find effects in the case of non-verified parent users increase over time, from a weak effect (0.13) in Phase I to moderate effects (0.21) in Phase II and III. In contrast, no substantial increase in peer effect is observed in the case of verified parent tweets.
- In dyadic interactions, SD (thoughts) parent tweets were more likely to be replied to with SD in child tweetswhen parent tweets were from non-organizational (human) verified users. We find this relationship for human verified accounts to have significant but weak effect size for all three phases (0.15, 0.17, and 0.18 respectively). For other dimensions of self-disclosure, peer effects across verified user types were negligible although significant. In



(a)

(b)

(c)

Figure 2: Self-disclosure in direct dyadic response to self-disclosing parent tweet by verified and non-verified users across the following categories: (a) Information; (b) Thoughts; (c) Feelings. x-axis: Proportion of SD/non-SD replies to SD/non-SD parent tweet; y-axis: SD in parent tweet (yes/no); orange = SD, blue = no SD.

contrast, we find that in longer conversations (3+ tweets) peer effects for thoughts disclosure were stronger for organizational accounts in the first phase. However, the effect was not sustained in subsequent phases. We observe similar trends for feelings.

The presence of peer effects on SD across categories of information sharing is corroborative of the literature exploring why and how individuals share on these platforms, noting relationship building as central, e.g., social penetration theory. Disentangling these effects for conversations started by verified vs unverified users suggests the unique role of these notable conversations in promoting, in particular, disclosure of thoughts. This is consistent with colloquial notions of public figures as opinion drivers and thought leaders.

We identified 15 topics dominating conversations across three phases, as described in Methods (see Figure 1). The most frequently discussed topics in the early days of the pandemic revolve around quarantine, President Donald Trump's words on the coronavirus, and the death rate. Topics about wearing masks and covid-19 testing become the center of discussion in the middle stages. In the third phase, new central discussion points around vaccines and school education emerge. Wearing masks remains a hot topic over three phases, and the number of tweets around it doubles in phase 2. Topics linked to vaccines and school education see exponential growth over time.

We zoom in on three important topics, selected because of their relevance in terms of size among the ones we identified and in terms of relevance for the pandemic narrative in Table 4. All three consistently result in higher degree of SD compared to general rate across the dataset. Among verified accounts, mask wearing is a popular topic throughout, for both humans and organizations. A growth trend is instead reported for unverified accounts where we note a jump of over 20% from phase I to phase II. Topics around schools and vaccines became more popular and personal for both account types over time, above the general rate of disclosure (about 5-10% consistent rate of SD).

Next, we consider the proportion of verified and unverified accounts involved in each topic, regardless of SD. Unverified accounts are highly engaged in all of the topics, and this trend is confirmed even as time passes. In phase I, verified accounts are generally more active in topics with keywords related to geo-political events (keywords -not reported for space include ”India, Pakistan, Indian Students” and topic 6 ”Iran, Russia, Iranian, UAE”), and much less on highly controversial topics (keywords: ”Trump, Hoax, new hoax, president”). We confirm this trend in phase II and in phase III, where high profile and controversial topics are least popular among verified users, and are mostly covered by unverified users. We note topics in phase II related to protests and riots, and virus spread and president, respectively. Also consistent is the heavy involvement in geo-political topics for verified accounts.

Conclusion

In this study, we present a comparison of self-disclosure in verified vs unverified users in Twitter during the Covid-19 pandemic. We discuss evidence of peer effects and topics of self-disclosing tweets. We find that verified accounts play a role in triggering and supporting conversations linked to health guidelines despite being significantly outnumbered by unverified users.

		Masks	**School**	**Vaccine**	**Total**
P1	*verif*	56	26	47	38
	human	72	47	62	58
	org	45	17	39	27
	unverif	74	71	71	71
P2	*verif*	65	51	53	48
	human	81	78	73	68
	org	52	40	44	35
	unverif	94	93	94	85
P3	*verif*	63	48	53	48
	human	82	67	73	69
	org	50	37	44	35
	unverif	93	91	92	85

Table 4: SD rate in conversations about selected topics by Phase(P). Total reflects all conversations, regardless of topic.

References

Campello, R. J. G. B.; Moulavi, D.; and Sander, J. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In Pei, J.; Tseng, V. S.; Cao, L.; Motoda, H.; and Xu, G., eds., *Advances in Knowledge Discovery and Data Mining*, 160–172. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. 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.

Lookadoo, K.; Hubbard, C.; Nisbett, G.; and Wong, N. 2021. We're all in this together: celebrity influencer disclosures about COVID-19. *Atlantic Journal of Communication*, 0(0): 1–22.

McInnes, L.; Healy, J.; and Melville, J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wang, Y.-C.; Burke, M.; and Kraut, R. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, 74–85.

Wang, Z.; Hale, S.; Adelani, D. I.; Grabowicz, P.; Hartman, T.; Flöck, F.; and Jurgens, D. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, 2056–2067. ACM.