Evolution of Intent and Social Influence Networks and their Significance in Detecting COVID-19 Disinformation Actors on Social Media ⋆ ⋆⋆⋆⋆⋆

Chathika Gunaratne[1], Debraj De[1], Gautam Thakur[1], Chathurani Senevirathna[2], William Rand[3], Martin Smyth[4], and Monica Lipscomb[4]

[1] Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA
{gunaratnecs, ded1, thakurg}@ornl.gov
[2] University of Central Florida
[3] North Carolina State University
[4] U.S. National Geospatial-Intelligence Agency

Abstract. Online disinformation actors are those individuals or bots who disseminate false or misleading information over social media, with the intent to sway public opinion in the information domain towards harmful social outcomes. Quantification of the degree to which users post or respond intentionally versus under social influence, remains a challenge, as individuals or organizations operating the profile are foreshadowed by their online persona. However, social influence has been shown to be measurable in the paradigm of information theory. In this paper, we introduce an information theoretic measure to quantify social media user intent, and then investigate the corroboration of intent with evolution of the social network and detection of disinformation actors related to COVID-19 discussions on Twitter. Our measurement of user intent utilizes an existing time series analysis technique for estimation of social influence using transfer entropy among the considered users. We have analyzed 4.7 million tweets originating from several countries of interest, during a 5 month period when the arrival of the first dose of COVID vaccinations were announced. Our key findings include evidence that: (i) a significant correspondence between intent and social influence; (ii) ranking over users by intent and social influence is unstable over time with evidence of shifts in the hierarchical structure; and (iii) both user intent and social influence are important when distinguishing disinformation actors from non-disinformation actors.

Keywords: Disinformation · Misinformation · COVID-19 · Intent · Social Influence · Twitter · Transfer Entropy · Information Theory

⋆ Supported by the U.S. National Geospatial-Intelligence Agency (NGA).
⋆⋆ Thanks to Cody Buntain of University of Maryland for supplying the Twitter dataset.

1 Introduction

Quantitative study into the role of user *intent* in online social dynamics is not yet well documented in the scientific literature. Nevertheless, intent is an important feature to consider when distinguishing regular users and misinformation actors from disinformation actors. Disinformation has been defined as the intentional dissemination of false or misleading information by malicious actors with the intent of swaying public opinion towards socially dangerous outcomes [6, 24, 25]. Therefore, by definition, without the measurement of intent, it is impossible to distinguish those instances of disinformation from misinformation. In this paper, we present a novel, information theoretic approach for estimating user intent on social networks. We applied this technique to analyze user intent and social influence expressed in COVID-19 discussions on Twitter, during the period 1st January 2021 to 21stMay 2021. We chose this time period as it was when the first doses of COVID-19 vaccinations started to be discussed online [20, 22]. In particular, we investigated whether there was a correspondence between user intent and social influence, and whether the ranking of users by intent and social influence remained stable over time. Users with high intent are likely to express their own agendas, acting as gate-keepers of new information and ideas into a social network, and users with high social influence have a stronger impact in swaying the opinions and behaviors of the other users in the network. Therefore, users with both high intent and high social influence were identified as those, who pose the highest risk of disinformation propagation, if their motivations were to be malicious or manipulative. Establishing whether the agendas of the users are malicious or manipulative is beyond the scope of this paper.

2 Background

During the COVID-19 pandemic, both misinformation and disinformation have played a major role in spreading confusion, fear, insecurity, and anti-public health narratives among targeted populations [19, 5]. Certain properties of disinformation help distinguished it from misinformation. While misinformation constitutes a claim that contradicts or distorts common understandings of verifiable facts [6], disinformation refers to such falsehoods that are *intentionally* propagated to actively undermine integrity in the information domain [6, 21]. In other words, disinformation may be distinguished by the intentional purpose to deceive, while misinformation may simply be a result of inadvertent or unintentional action [6]. Thus, *intent or intentionality is the major discriminator between misinformation and disinformation* [24]. Proving intent in users or accounts can sometimes be more challenging than just identifying falsehoods in content [6]. Detecting intent is hard because of the difficulty to uncover ground truth beliefs in people/accounts about the veracity of information content, and the further difficulty in ascertaining their underlying motivations [25]. The current literature states that recognizing the range of motivations for spreading misinformation is valuable, even if the motivations or intentions are hard to disentangle [25]. This is the key motivation behind this study, attempting to quantitatively measure intent and analyze the dynamics of intent and social influence of social media accounts through time.

Most of the existing literature regarding social influence mainly utilized social network centrality, link-topology, and coreness-based measurements to quantify social influence [1, 13, 9, 26]. However, these techinques depend on the underlying network structures of user connectivity, which in turn are typically constructed using the follower-followee network (such as in Twitter) or friendship network (such as in Facebook). But, follower-followee networks or friendship networks represent the users' popularity, and it has been shown that the relation between structural influence and users activities is weak [3]. In addition to these techniques, some studies have used entropy-based measures, which were based upon network structure [12, 4] or an information-theoretic approach [23, 2, 7, 8, 17]. In this regard, we have utilized the quantification of social influence from our previous work [7, 8, 17], to calculate *social influence* in this work in order to infer *user intent*.

3 Methodology

We introduce a novel information theoretic approach to the quantification of intent from social media user activity timeseries data, which we applied to a COVID-19 Twitter dataset. We then used two machine learning approaches to classify users as disinformation or non-disinformation actors and used the generated labels to study the importance of social influence and intent on disinformation actor detection.

3.1 Data

We analyzed a dataset of 4,714,617 tweets on the COVID-19 pandemic between January 1st 2021 and May 21st 2021. This data consisted of 14,876 unique users with at least 10 actions (tweets, replies, retweets, and quoted tweets) per month, to ensure meaningful statistical results. The Twitter data was collected as follows. From the GeoCov19 dataset [14], we identified user accounts that have inferred profile- and message- based locations in few countries of interest (Australia, Brazil, Canada, Britain, India, Nigeria, New Zealand, Taiwan, South Africa). Then for these users we collected their tweets, replies, retweets, and quoted tweets, along with those by other users that responded to this activity with replies and retweets, during the time period considered.

3.2 An Information Theoretic Approach to Intent Measurement

We expand on the information theoretic measurement of social influence introduced in our previous studies [7, 8, 17], which show that given the activity time series of a set of online social media users (say V), the social influence experienced by a user of interest, $u \in V$, due to another user, $v \in V$, can be measured using transfer entropy (say $T_{v \to u}$). $T_{v \to u}$ is defined in the equation eq. 1, where t is the current time step, T the entire time period analyzed, and k is history length. In this study, we consider a time step as 1 week, and $k = 1$. Transfer entropy is a directional measure of the information transfer between two random processes. In the case of social networks, it can be utilized to measure the information transfer from the activity time series of v to that of u, acting as an estimator

of social influence. If $T_{v \to u} > 0$ a social influence link exists between the two users and v has a certain magnitude of influence over u.

$$T_{v \to u} = \sum_{t \in T} P(u_t, u_{t-1:t-k}, v_{t-1:t-k}) log \frac{P(u_t \mid u_{t-1:t-k}, v_{t-1:t-k})}{P(u_t \mid u_{t-1:t-k})} \tag{1}$$

In this study, we utilized transfer entropy-based estimation of social influence to measure the degree of intent with which users performed actions online. The Shannon entropy of u, H_u, measures the overall information produced by activity of u. We premise that, given sufficient sources of the social influences on u, the information intentionally produced by u would be the Shannon entropy of u minus the sum of all transfer entropy experienced by u, as shown in eq. 2.

$$I_u = H_u - \sum_{v \in V} T_{v \to u} \tag{2}$$

Similarly, we computed the total influence exerted by the user of interest (u), say T'_u, as the total transfer entropy exerted by a user u on all other users considered, as shown in eq. 3.

$$T'_u = \sum_{v \in V} T_{u \to v} \tag{3}$$

These two measurements: (ii) degree of user intent (I_u) and (ii) total social influence exerted (T'_u), were used in our analysis to better understand the social influence and user intent dynamics of COVID-19 related discussions on Twitter and disinformation actor detection. Tweets, replies, retweets, and quoted tweets were considered as user actions and the respective action timeseries for each user in the collected data was reconstructed. We test the following *three hypotheses* using these two measurements:

- Hypothesis I: There is a significant correspondence between high intent and high social influence.
- Hypothesis II: The ranking of users by intent and by social influence remains stable over time, i.e. users with high intent and high influence remain so, and vice versa.
- Hypothesis III: There is a significant difference in user intent among disinformation actors from that among non-disinformation actors.

3.3 Disinformation Classification

We constructed two models to classify each user as either a disinformation actor (i.e., IO - information operative) or a non-disinformation actor (i.e., Real). Specifically, we utilized: (1) a weakly-supervised classification model based on Snorkel [15]; and (2) a logistic regression model.

The weakly-supervised model used Snorkel's labeling function system to encode human cognitive heuristics and fit a weight matrix of conditional probabilities of outputting a particular label. This was based on the label votes of a set of labeling functions provided during training. We used Snorkel labeling functions implemented for detection of IO on Twitter from recent literature [18]. The Snorkel label model classified each user

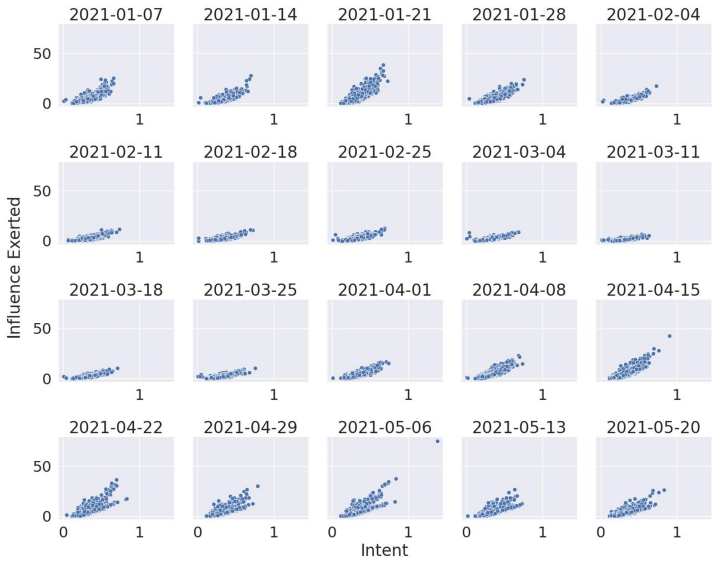


Fig. 1: Distribution of total social influence exerted by users vs intent, by weekly passage of time between Jan 1st and May 21st in 2021. Weeks progress in ascending order from left to right, and from top to bottom.

as: *IO*, or *Real*, or *Undecided* (in the case of a tied vote). We replaced Undecided users (approximately 3.65% of the predicted Snorkel labels), with a uniform random choice between *IO* and *Real*, to avoid bias towards either class.

The logistic regression model was trained on features engineered on the collected data. We generated a suite of 32 features, belonging to six broad categories as follows: (i) user social influence and intent; (ii) tweet statistics on emoji, hashtag, mention, character count, etc.; (iii) temporal tweets characteristics; (iv) user profile characteristics; (v) tweets ratio characteristics; (vi) other characteristics like tweet count, date range, etc. Relevant features were selected after an extensive review of existing literature, and also exploratory data analysis on disinformation dataset released by Twitter's Information Operations group [10]. A detailed discussion of the considered features is beyond the of scope of this paper. The logistic regression model was trained to label users as *IO* or *Real*, based on the engineered features, and Snorkel labels (from first model) were used as ground truth for training. We found that the regression model fit the Snorkel labeled data reasonably well (precision = 0.87, recall = 0.86, f1-score = 0.86).

4 Results

We tested Hypothesis I, by examining the correlation between intent and total social influence exerted over time, as shown in Figure 1. A distinct correlation was seen between intent and total social influence exerted. Furthermore, we observed that the relationship between intent and total social influence exerted, changed with the progression of time (through weeks). Specifically, from 2021-02-04 till 2021-03-25, social influence exerted was strikingly lower even for high intent individuals. Overall, a Pearson correlation test

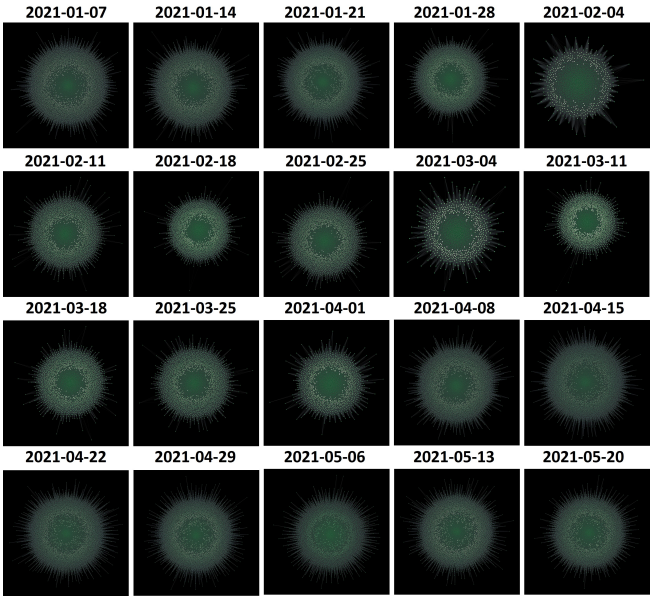
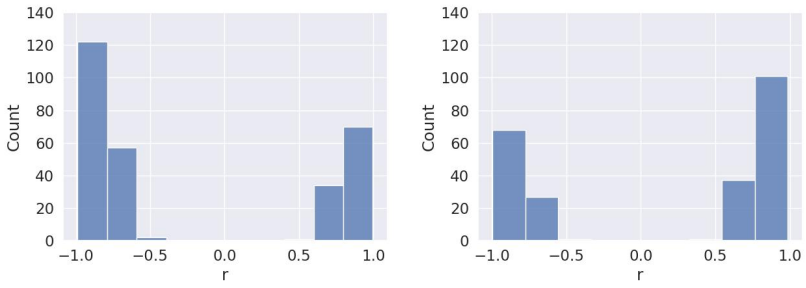


Fig. 2: Visualization of social influence and user intent networks over the passage of time (weeks) between Jan 1st and May 20th. Low intent individuals are colored darker green, higher intent individuals are brighter yellow. Individuals towards the center of the networks have higher connectivity (social influence), and users towards the outer part of the networks have lower connectivity. Weeks progress in ascending order from left to right, and from top to bottom.

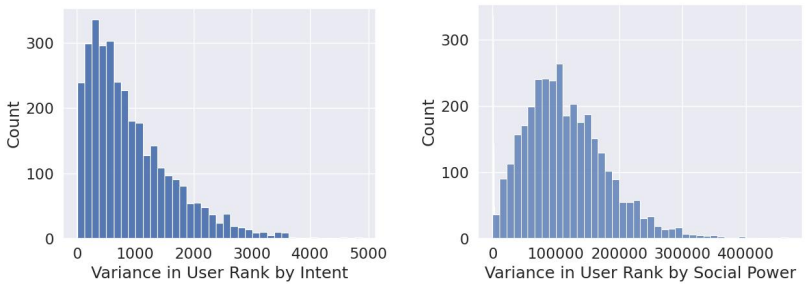
revealed a correlation coefficient of $r = 0.6961$ and $p \approx 0$. Additionally, Figure 2 displays the snapshots of the social network captured over the progression of time (weeks). Nodes' color intensities signify higher intent, and nodes with higher social influence links are towards the center of the network. The brightly colored ring towards the middle of each social network indicates individuals of high intent. There was a slight shift of this ring towards the center of the network starting at week 6 (2021-02-11) until week 13 (2021-03-18). Users with both high intent and high social influence pose high risk for spread of disinformation and likely existed within this band.

To test Hypothesis II, we performed Pearson's correlation tests on both intent and total social influence exerted over time. 3,306 users with at least 5 weeks of activity were tested. A significance level of $\alpha = 0.05$ was assumed and users with $p > \alpha$ were not considered. Figure 3 displays the correlation coefficients of both intent and total social influence exerted among the user population. A strong bi-modality, with many users either having strong positive correlations or negative correlations for both intent and total social influence exerted was observed. However, it is important to note that only 235 out of the 3,306 users had a $p < 0.05$, meaning the rest of the users had insufficient data to produce sufficient confidence in the Pearson correlation test. Within this set of users we find evidence against Hypothesis II, showing that there can indeed be considerable shift in both intent and total social influence exerted over time within the social network.



(a) (b)

Fig. 3: Distribution of users' Pearson correlation coefficients (r) of: (a) intent and (b) total social influence exerted with time.



(a) (b)

Fig. 4: Distributions of variance in user ranking by: (a) intent and (b) total social influence exerted rank, over a 20 week period.

Furthermore, we measured the change in rank of users based on their intent and total social influence exerted. As shown in Figure 4, we observed a difference in variance in ranking of users over time changes when considering intent versus total social influence exerted. Particularly, users had much greater variance in ranking by total social influence exerted, than ranking by intent, suggesting that it was more common to see changes in the social network hierarchy, than it was to see changes in ranking by intent. Additionally, the distribution of variance in user rank by intent was highly-skewed, in contrast to that of variance in rank by total social influence exerted, which indicated that while large changes in ranking by total social influence exerted among users may be more normal among the population, it was less common for users to change their ranking by intent.

Finally, in order to test *Hypothesis III*, we examined the correspondence of user intent and total social influence exerted with disinformation actors, as identified by the Snorkel labeling heuristics model and the regression classifier model (both models were described in Section 3.3). Figure 5 compares the degree of intent of disinformation actors versus that of non-disinformation actors as classified by the Snorkel heuristics, and Figure 6 displays the same comparison for total social influence exerted. By conducting Mann-Whitney U tests at 95% confidence, we found support for the alternate hy-

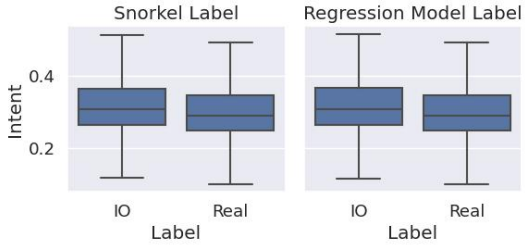


Fig. 5: Comparison the user intent of disinformation actors vs non-disinformation actors as predicted by the Snorkel weak-supervision model.

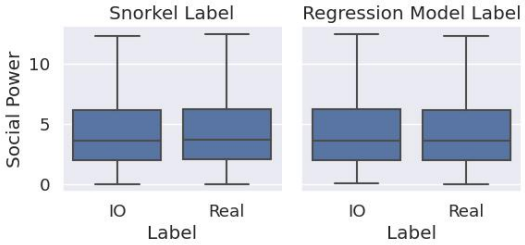


Fig. 6: Comparison total social influence exerted of disinformation actors vs non-disinformation actors as predicted by the Snorkel weak-supervision model.

pothesis that intent of non-disinformation actors was significantly less than that of disinformation actors, as classified by both the models: Snorkel weak supervision labels ($U = 408319163.5$, $p = 1.7397 \times 10^{-143} < 0.05$), and regression classifier labels ($U = 428923903.0$, $p = 1.7397 \times 10^{-143} < 0.05$). However, Mann-Whitney U tests at 95% confidence, for the alternate hypothesis that social influence exerted by non-disinformation actors was less than that of disinformation actors, were not supported for both the Snorkel weak supervision labels ($U = 361421379.0$, $p = 0.9990 > 0.05$) and the regression classifier labels ($U = 382374824.5$, $p = 0.2641 > 0.05$).

5 Discussions and Conclusions

In this paper, we introduce a novel entropy-based approach to measure user intent towards posting in online social networks using an entropy-based method. We used this technique to analyze the dynamics of intent and the evolution of social influence on a network of Twitter users discussing COVID-19. The use of our proposed measures for user intent and total social influence exerted led to several interesting and novel findings as elaborated below.

We found that there was a significant correspondence between intent and total social influence exerted, and this relationship changes over time. As shown in Figure 1, inside the 20 consecutive weeks of analysis, the relationship between the Influence Exerted and the Intent (the slope of a regression line from the scattered data points) remained relatively strong from the week of 2021-01-07 for 4 weeks, after which it remained weak from 2021-02-04 till the week of 2021-03-25. Then the relationship grew again and remained at its initial strength throughout the remaining 9 weeks of our analysis period.

This was likely due to an exogenous shock to the influence network during this period. Interestingly, we have observed that during that 8 week period the intensity of news regarding COVID-19 vaccine emergency authorizations and mobilization of vaccine roll-outs by the United States Food and Drug Administration (FDA) and World Health Organization (WHO) heightened greatly [22][20]. Conversely, it was towards the end of the 8 week period when vaccination rates gain momentum for the global low-income population [11][16]. Overall, it seems that when news of mobilization in vaccine deliveries were initiated, users with higher intent lost some degree of the social influence they exerted. But when global low-income population's vaccination gained momentum, users with higher intent likely resumed exerting more social influence like before (before the news of vaccine deliveries started).

We found that the ranking by intent and social influence evolved significantly over time at the microscopic scale, while the distributions remained relatively stable at the macro-scale. Our findings contradicted Hypothesis II, providing evidence that ranking of users by intent and social influence was not necessarily stable over time. We found evidence that a reasonable portion of individuals have high variance in rank by both intent and social influence. Furthermore, we observed that a significant number of individuals have either strong positive or strong negative shifts in intent and social influence over time. This indicated that there was a reasonable amount of evolution in the social hierarchy of the considered population over time.

Most importantly, we found that there was a statistically significant increase in intent among disinformation actors, in comparison to that of non-disinformation actors. This partially supported Hypothesis III, such that disinformation actors can be distinguished by the degree of intent in their activity. Conversely, we found evidence that total social influence exerted may be similar for both disinformation and non-disinformation actors, likely reducing its importance when identifying disinformation actors.

Overall, our findings help further the state-of-the-art in understanding disinformation dynamics and evolution of online social networks. We have shown that intent of user activity has a significant impact on online information dynamics, and is an important feature to be included in the detection of disinformation actors.

References

1. Al-Garadi, M.A., Varathan, K.D., Ravana, S.D., Ahmed, E., Mujtaba, G., Khan, M.U.S., Khan, S.U.: Analysis of online social network connections for identification of influential users: Survey and open research issues. ACM Comput. Surv. **51**(1) (Jan 2018)
2. Bhattacharjee, A.: Measuring influence across social media platforms: Empirical analysis using symbolic transfer entropy (2019), https://scholarcommons.usf.edu/etd/7745
3. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.: Measuring user influence in twitter: The million follower fallacy. Proceedings of the International AAAI Conference on Web and Social Media **4**(1) (May 2010), https://ojs.aaai.org/index.php/ICWSM/article/view/14033
4. Chen, X., Zhou, J., Liao, Z., Liu, S., Zhang, Y.: A novel method to rank influential nodes in complex networks based on tsallis entropy. Entropy **22**(8), 848 (Jul 2020)
5. Gottlieb, M., Dyer, S.: Information and disinformation: social media in the covid-19 crisis. Academic emergency medicine (2020)
6. Guess, A.M., Lyons, B.A.: Misinformation, disinformation, and online propaganda. Social media and democracy: The state of the field, prospects for reform pp. 10–33 (2020)

7. Gunaratne, C., Baral, N., Rand, W., Garibay, I., Jayalath, C., Senevirathna, C.: The effects of information overload on online conversation dynamics. Computational and Mathematical Organization Theory **26**(2), 255–276 (Jun 2020)
8. Gunaratne, C., Rand, W., Garibay, I.: Inferring mechanisms of response prioritization on social media under information overload. Scientific reports **11**(1), 1–12 (2021)
9. Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identification of influential spreaders in complex networks. Nature Physics **6**(11), 888–893 (2010). https://doi.org/10.1038/nphys1746
10. Operation, T.I.: Insights into attempts to manipulate Twitter by state linked entities. (2022), https://transparency.twitter.com/en/reports/information-operations.html
11. Our World in Data: Global Coronavirus (COVID-19) vaccinations dashboard. (2021), https://ourworldindata.org/grapher/cumulative-covid-vaccinations-income-group
12. Peng, S., Li, J., Yang, A.: Entropy-based social influence evaluation in mobile social networks. pp. 637–647. Springer, Cham (Nov 2015). https://doi.org/10.1007/978-3-319-27119-4_44
13. Peng, S., Zhou, Y., Cao, L., Yu, S., Niu, J., Jia, W.: Influence analysis in social networks: A survey. Journal of Network and Computer Applications **106**, 17 – 32 (2018)
14. Qazi, U., Imran, M., Ofli, F.: Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information (2020)
15. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: rapid training data creation with weak supervision. The VLDB Journal **29**(2), 709–730 (May 2020)
16. Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., Roser, M.: Coronavirus pandemic (covid-19). Our World in Data (2020), https://ourworldindata.org/coronavirus
17. Senevirathna, C., Gunaratne, C., Rand, W., Jayalath, C., Garibay, I.: Influence cascades: Entropy-based characterization of behavioral influence patterns in social media. Entropy **23**(2), 160 (2021)
18. Smith, S.T., Kao, E.K., Mackin, E.D., Shah, D.C., Simek, O., Rubin, D.B.: Automatic detection of influential actors in disinformation networks. Proceedings of the National Academy of Sciences **118**(4) (2021)
19. Tagliabue, F., Galassi, L., Mariani, P.: The "pandemic" of disinformation in covid-19. SN comprehensive clinical medicine **2**(9), 1287–1289 (2020)
20. The American Journal of Managed Care: A Timeline of COVID-19 Vaccine Developments in 2021. (2021), https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021
21. Tucker, J.A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B.: Social media, political polarization, and political disinformation: A review of the scientific literature. Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018) (2018)
22. U.S. DoD: U.S. DoD Coronavirus Timeline. (2021), https://www.defense.gov/Spotlights/Coronavirus-DOD-Response/Timeline/
23. Ver Steeg, G., Galstyan, A.: Information transfer in social media. In: Proceedings of the 21st International Conference on World Wide Web. p. 509–518. WWW '12, Association for Computing Machinery, New York, NY, USA (2012)
24. Wardle, C., et al.: Information disorder: The essential glossary. Harvard, MA: Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School (2018)
25. Wittenberg, C., Berinsky, A.J.: Misinformation and its correction. Social Media and Democracy: The State of the Field, Prospects for Reform **163** (2020)
26. Zeng, A., Zhang, C.J.: Ranking spreaders by decomposing complex networks. Physics Letters A **377**(14), 1031–1035 (2013)