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], and William Rand[3]

[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

Abstract. Online disinformation actors are those individuals or bots who spread false or misleading information on social media, with 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 on the social media, remains a challenge, as individual or organization operating the profile is 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 influence network of COVID-19 related discussions on Twitter. Our measure of user intent utilizes an existing time series analysis technique for estimation of social influence using transfer entropy along with total entropy measurement of the considered users. We have analyzed 4.7 million tweets globally (from several countries of interest) during almost 5 months period of interest (online discourse during arrival of first dose of COVID vaccination). Our analysis results have multiple key findings: (i) there is 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 National Geospatial-Intelligence Agency (NGA).
⋆⋆ Thanks to Cody Buntain of University of Maryland for supplying the Twitter dataset.

1 Introduction

The role of *intent* in online social dynamics is not yet well understood in the scientific literature. However, intent is an important feature to consider when distinguishing regular users and misinformation actors, from the 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 [8, 24, 25]. Therefore, by definition, without measurement of intent, it is impossible to distinguish those instances of misinformation from disinformation. ■ In this paper we present a novel, information theoretic approach for estimating user intent on social networks. We then use 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 [13, 6]. In particular, we investigate whether there is a correspondence between user intent and social influence, and whether the ranking of users by intent and social influence remain 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 effect on the opinions and behaviors of the other users in the network. Therefore, users with both high intent and high social influence are 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 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 [21, 7]. Certain properties of disinformation help distinguished it from misinformation. While misinformation constitutes a claim that contradicts or distorts common understandings of verifiable facts [8], disinformation refers to such falsehoods that are deliberately or *intentionally* propagated to actively undermine integrity in the information domain [8, 22]. In particular, disinformation is distinguished by the intentional purpose to deceive, while misinformation may simply be a result of inadvertent or unintentional action [8]. 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 [8]. Furthermore, detecting intent is hard because of the difficulty to uncover ground truth beliefs in people/accounts about the veracity of an information content, and even more difficulty in ascertaining their underlying motivations [25]. The research 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*, which is to quantitatively measure intent, and then analyze the dynamics of intent and social influence of accounts through time (through different weeks during certain phase of COVID-19 pandemic).

Regarding social influence, most of the prior works mainly used social network centrality, link-topological, and coreness-based measurements to quantify the social influence [1, 16, 12, 26]. However, these measurements depend on the underlying network structures of users, which were 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 work in [3] showed that relation between the structural influence and the user activities is weak. In addition to these measurements, some works have used entropy-based measures, which were based upon network structure [15, 4] or an information-theoretic approach [23, 2, 9, 10, 19]. In this regard, we have utilized the quantification of social influence from our previous work [9, 10, 19], to calculate *social influence* in this work in order to infer *user intent*.

3 Methodology

We introduce a novel information theoretic approach to quantification of user intent from rates of user activity over time.

3.1 Data

We analyze 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 events per month, to ensure meaningful statistical results. The overall global Twitter dataset was collected as follows. From the original GeoCov19 dataset [17], we identified user accounts who 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 and also related users' tweets during the time period of our interest. Related users are the users involved in replies and retweets. To note that in this process, the related users also belonged to several countries outside of our initial countries of interest.

3.2 An Information Theoretic Approach to Intent Measurement

We expand on the information theoretic measurement of social influence introduced in our previous work [9, 10, 19]. Given the activity time series of a set of online social media users (say V), these studies have shown that 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 activity time series of v to that of u, acting as an estimator for 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.

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

Meanwhile, the Shannon entropy of u, H_u, measures the overall information produced by activity of u. Our premise is that, given sufficient sources of the social influencers of u, the information intentionally produced by u would be the Shannon entropy of u minus the sum of all transfer entropies to u, as shown in eq. 2.

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

Similarly, we compute 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.

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

We use these *two measurements*: (i) T'_u for total social influence exerted; and (ii) I_u for degree of user intent. T'_u and I_u are used throughout our analysis in this work to better understand the social influence and user intent dynamics of COVID-19 related discussions on Twitter. 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 the non-disinformation actors.

3.3 Disinformation Classification

We have construct two models to classify each user as either disinformation actor (i.e., IO - information operative) or non-disinformation actor (i.e., Real). Specifically, we utilize: (1) a weakly-supervised classification model based on Snorkel [18]; and (2) a logistic regression model. ■ In *first* model, Snorkel uses a labeling function system to encode human cognitive heuristic and fits a weight matrix of conditional probabilities of outputting a particular label. This is based on the label votes of a set of labeling functions provided during training. We use Snorkel labeling functions implemented for detection of IO on Twitter from recent literature [20]. The Snorkel label model classifies each user as: *IO*, or *Real*, or *Undecided* (in the case of a tied vote). For Undecided users, we replaced with a uniform random choice between *IO* and *Real*. ■ In *second* model, we create and train a logistic regression model with focus on feature engineering. We generate a suite of 32 features, overall belonging to six 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;

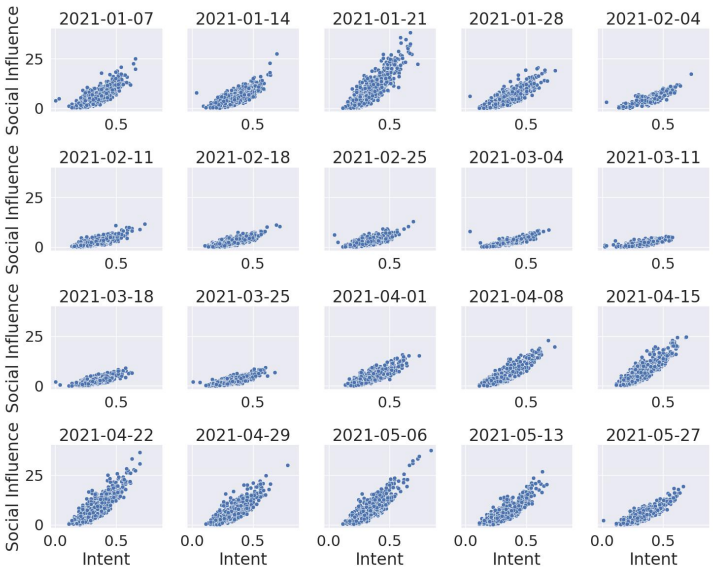


Fig. 1: Distribution of total social influence exerted by users vs intent, by weekly passage of time between Jan 1st and May 21st in 2021. Weeks progress in ascending order from left to right, and from top to bottom.

(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 [14]. A detailed discussion of the considered features is beyond the of scope of this paper. The logistic regression model is trained to label users as *IO* or *Real*, based on the engineered features, and Snorkel labels (from first model) are 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 and Analysis

■ *Hypothesis I* helps us obtain a macro-scale perspective of the relationship between intent and social influence. We test Hypothesis I, by examining the correlation between intent and total social influence exerted over time as shown in Figure 1. A distinct correlation can be seen between intent and total social influence. Furthermore, we see that the relationship between intent and total social influence exerted changes with the progression of time (through weeks). Specifically, from 2021-02-04 till 2021-03-25, social influence exerted is strikingly lower even for high intent individuals. Overall, a Pearson correlation test revealed a correlation coefficient of $r = 0.6961$ and $p \approx 0$. Additionally, Figure 2 displays the snapshots of the social network over the progression of time (weeks). Nodes' color intensity signifies higher intent, and nodes with higher social influence links are towards the center of the network. It can be seen that there is a ring of high intent individuals towards the middle of the social network. There is a slight shift

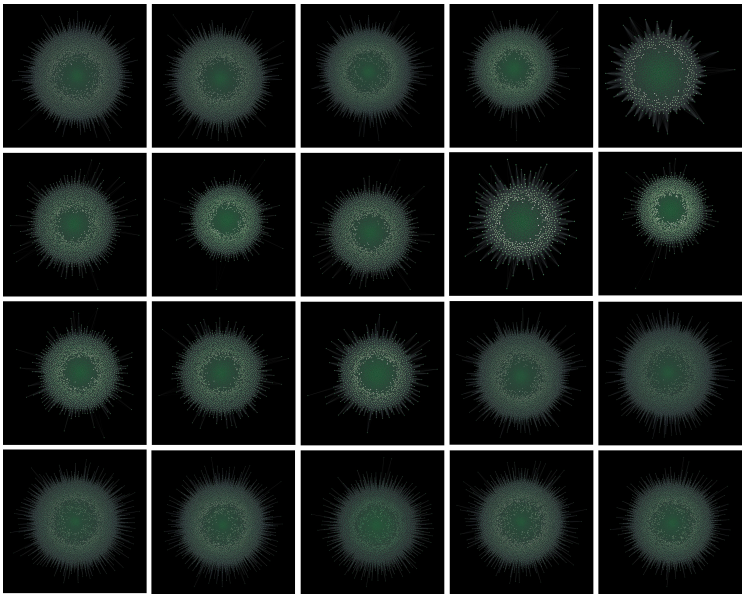
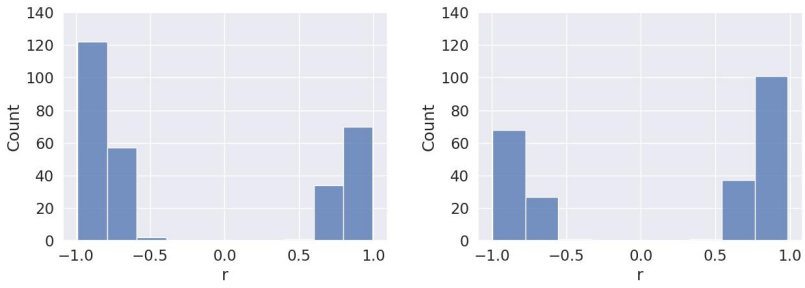


Fig. 2: Visualization of social influence and user intent networks over the passage of time (weeks) between Jan 1st and May 21st. Low intent individuals are colored darker green, higher intent individuals are brighter yellow. Individuals towards the center of the networks have higher connectivity (social 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.

of this ring towards the center of the network starting at week 6 (2021-02-11) till week 13 (2021-03-18). Users with both high intent and high social influence are considered high risk for spread of disinformation and likely exist within this band.

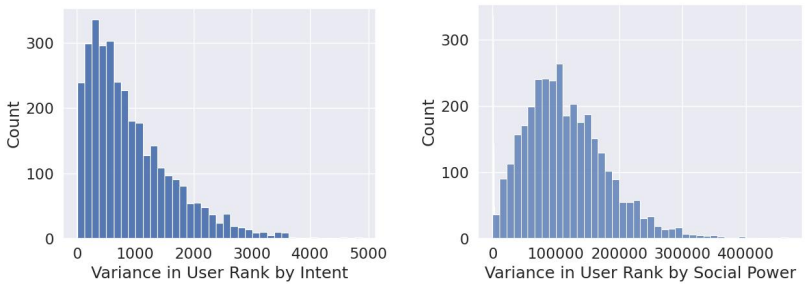
■ In order to test *Hypothesis II* we performed Pearson's correlation tests on both intent and total social influence exerted by time. Total 3,306 users with at least 5 weeks of activity were tested. A significance level of $\alpha = 0.05$ was assumed and users with $p > \alpha$ were not considered. Figure 3 displays the correlation coefficients of both intent and total social influence exerted among the user population. It can be seen that there is a strong bi-modality, with many users either having strong positive correlations or negative correlations for both intent and total social influence exerted. 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 the 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.

Furthermore, we look at the change in rank of users based on their intent and total social influence exerted. As shown in Figure 4, we observe a difference in variance in ranking of users over time changes when considering intent versus total social influence exerted. Particularly, users have much greater variance in ranking by total social influence exerted than the same by intent, suggesting that it is more common to see changes in



(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.

the social network hierarchy, than it is to see changes in ranking by intent. Additionally, the distribution of variance in user rank by intent is highly-skewed, in contrast to that of variance in rank by total social influence exerted, indicating that while large changes in total social influence exerted ranking among users may be more normal among the population, it is less common for users to change their ranking by intent.

■ Finally, in order to address *Hypothesis III*, we examine 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 performs the same comparison for total social influence exerted. By conducting Mann-Whitney U tests at 95% confidence, we found support for the alternate hypothesis that intent of non-disinformation actors was 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

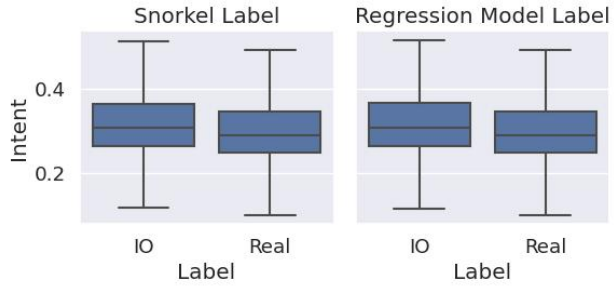


Fig. 5: Comparison the user intent of disinformation actors vs non-disinformation actors as predicted by the Snorkel weak-supervision model.

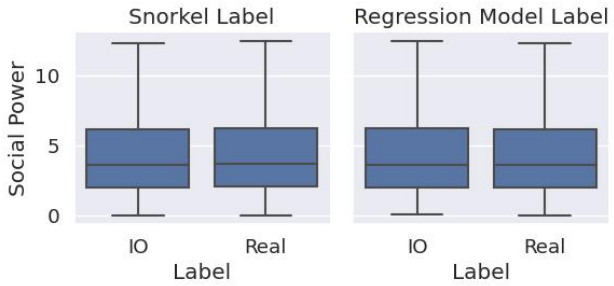


Fig. 6: Comparison total social influence exerted of disinformation actors vs non-disinformation actors as predicted by the Snorkel weak-supervision model.

less than that of disinformation actors, was 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 work we introduce a novel entropy-based approach to measure user intent towards posting in online social networks using an entropy-based method. We use our 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 has led to several interesting and novel findings as elaborated below.

■ We find 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 is 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 [6][13]. Conversely, it is towards the end of the 8 week period when vaccination rates gain momentum for the global low-income population [5][11]. 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 find that the ranking by intent and social influence evolves significantly over time at the microscopic scale, while the distributions remain relatively stable at the macro-scale. Our findings contradict out Hypothesis II that ranking of users by intent and social influence is stable over time. We find evidence that a reasonable portion of individuals have high variance in rank by both intent and social influence. Furthermore, we find that a significant number of individuals have either strong positive or strong negative shifts in intent and social influence over time. This indicates that there is a reasonable amount of evolution in the social hierarchy of the considered population over time.

■ Importantly, we find that there is a statistically significant increase in intent among disinformation actors, in comparison to that of non-disinformation actors. This partially supports Hypothesis III, such that disinformation actors can be distinguished by the degree of intent in their activity. However, we find 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 analysis results and 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 particularly in identification 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. in Data, O.W.: Global Coronavirus (COVID-19) vaccinations dashboard. (2021), https://ourworldindata.org/grapher/cumulative-covid-vaccinations-income-group?country=High+income Low+income Lower+middle+income Upper+middle+income
6. DoD, U.: U.S. DoD Coronavirus Timeline. (2021), https://www.defense.gov/Spotlights/Coronavirus-DOD-Response/Timeline/

7. Gottlieb, M., Dyer, S.: Information and disinformation: social media in the covid-19 crisis. Academic emergency medicine (2020)
8. 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)
9. 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)
10. Gunaratne, C., Rand, W., Garibay, I.: Inferring mechanisms of response prioritization on social media under information overload. Scientific reports **11**(1), 1–12 (2021)
11. Hannah Ritchie, Edouard Mathieu, L.R.G.C.A.C.G.E.O.O.J.H.B.M.D.B., Roser, M.: Coronavirus pandemic (covid-19). Our World in Data (2020), https://ourworldindata.org/coronavirus
12. 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
13. of Managed Care, T.A.J.: A Timeline of COVID-19 Vaccine Developments in 2021. (2021), https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021
14. Operation, T.I.: Insights into attempts to manipulate Twitter by state linked entities. (2022), https://transparency.twitter.com/en/reports/information-operations.html
15. 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
16. 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)
17. Qazi, U., Imran, M., Ofli, F.: Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information (2020)
18. 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)
19. 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)
20. 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)
21. Tagliabue, F., Galassi, L., Mariani, P.: The "pandemic" of disinformation in covid-19. SN comprehensive clinical medicine **2**(9), 1287–1289 (2020)
22. 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)
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)