A preliminary investigation of COVID-19 transmission in the United States by incorporating social media sentiments

Tianyi Tan¹ | Teng Huang² | Xi Wang¹ | Zhiya Zuo³

¹School of Information, Central University of Finance and Economics, Beijing, China
²Department of Operations and Information Management, University of Connecticut, Mansfield, Connecticut
³Department of Information Systems, City University of Hong Kong, Hong Kong, China

Correspondence
Tianyi Tan, School of Information, Central University of Finance and Economics, Beijing, China.
Email: 2017312322@email.cufe.edu.cn

Abstract
COVID-19 has now become a global pandemic. During the widespread of COVID-19, Twitter, as an online social media platform, has been a preferred channel for interaction and communication. As a result, it provides huge amount of information from which latent signals such as sentiments can be mined for a better understanding of COVID-19 transmission patterns. As a preliminary attempt, we reveal a strongly positive zero-order correlation between sentiments of tweets and COVID-19 confirmed cases in U.S. Considering the unique hierarchical structure of the U.S. government, state governments exert their own power to issue public health policies. Indeed, there are different patterns of correlations between sentiments and COVID-19 confirmed cases, affirming that country-level characteristics suppress that of state-level. Diving deeper into the textual content of COVID-19 related tweets, there manifests a diverse set of topics which in turn lead to dispersed sentiments. Our preliminary investigation paves the way for a finer-grained analysis of the COVID-19 transmission and social media activities by considering varying situations across states and topics.

KEYWORDS
COVID-19, sentiment analysis, social media

1 INTRODUCTION

While China is the first country that suffers from the outbreak, the novel coronavirus has made its way to cause a global pandemic, as announced on March 11th 2020 by the World Health Organization (WHO, Forthcoming). The U.S., in particular, has reported over 82,000 confirmed cases on March 26th 2020, the highest number in the world (Smith, 2020). By June 2020 (CDC, 2020), this number has come to 1.8 million, along with a total number of over 100,000 death cases. In order to reduce social interaction, public health measures such as school and workplace closure have stimulated more communication via online social media platforms. In this regard, Twitter, one of the most popular social media platforms in the U.S., is an ideal venue through which subtle signals may be mined for better explanation and prediction of COVID-19 transmission patterns.

Using a publicly available Twitter dataset (Chen, Lerman, & Ferrara, 2020), we investigate the relationship between sentiments and COVID-19 transmission in the U.S. In total, we gathered over 5 million tweets posted by users whose residential addresses are in the U.S. from March 1st to April 25th. It is noteworthy that we
deliberately select the starting date to be in March, from which social distancing measures started to be adopted. When preprocessing the data of posting users’ self-reported locations in their Twitter user profiles, we exclude the non-US tweets so that we are able to label each tweet with 1 out of 50 states. Currently we leave out District of Columbia and Puerto Rico due to their subtly different natures with other states, which will be incorporated in the future. Sentiment scores from $-1$ (most negative) to 1 (most positive) are labelled for each tweet via VADER, a lexicon and ruled based tool specifically designed for social media (Hutto & Gilbert, Forthcoming).

2 | RESULTS

In our analysis, sentiments are widely spread with a standard deviation of 0.4 and a median of 0. Interestingly, there exists a strong correlation between the logarithm of daily confirmed cases and the average as well as standard deviation of sentiments from tweets posted on the same day (Figure 1).

It is noteworthy that the U.S. has a unique structure of government hierarchies. State governments retain strong powers including elections and public health measures. Given varying situations across various states, governors have taken different responses to contain COVID-19. For example, as the first state in the U.S. that announced a confirmed case in January, Washington declared the state of emergency on February 29th. However, face masks are only mandatory for employee in public-facing businesses effective from May 4th. New Jersey, meanwhile, imposed a more stringent measure requiring all individuals to use face masks in public spaces from April 8th. See Raifman et al. (2020) for a more comprehensive listing of state-wise policies regarding COVID-19 in the U.S. All these imply the existence of contextual differences across states, which in turn affects both the social media activities and COVID-19 spread patterns. We therefore hypothesize that the country-level analysis as depicted in Figure 1 inevitably suppresses state-level characteristics. Figure 2 confirms our hypothesis by visualizing the relationship between daily confirmed cases and social media sentiments at state level.

**FIGURE 1** Trends of daily confirmed cases (in logarithm) and sentiments of COVID-19 related tweets (average and standard deviation)
In closing, the preliminary results validate our hypotheses. The positive correlations between confirmed cases and both the central tendency and variation of sentiments, in particular, suggest an in-depth analysis at both topic and state level rather than tweet level. On closer inspection over the tweet contents, it is found that the increasing online communication due to social distancing composites a highly diverse group of topics, including celebrating recovery from COVID-19 (e.g., “...God is truly good and I'm happy my mom was able to beat covid 19...”), criticizing the government (e.g., “...Our imbecile president gave incoherent remarks about the US plan to manage coronavirus...”), grieving (e.g., “...We've lost more New Jerseyans to coronavirus than we lost in World War I, the Korean War, and the Vietnam War...”), etc. As a result, we plan to utilize aspect-based sentiment analysis for finer-grained portrait of public sentiment dynamics in social media which can in turn facilitate model constructions to explain the current situation of and predict future patterns of COVID-19 spread (Perikos & Hatzilygeroudis, 2017).

REFERENCES

CDC. (2020). Coronavirus Disease 2019 (COVID-19) in the U.S. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html
Chen, E., Lerman, K., & 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. https://doi.org/10.2196/19273

Hutto, C. J., & Gilbert, E. (Forthcoming). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text 10.

Perikos, I., & Hatzilygeroudis, I. (2017). Aspect based sentiment analysis in social media with classifier ensembles. *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, 273–278. https://doi.org/10.1109/ICIS.2017.7960005

Raifman, J., Nocha, K., Jones, D., Bor, J., Lipson, S., Jay, J., & Chan, P. (2020). COVID-19 US state policy database (CUSP). Retrieved from www.tinyurl.com/statepolicies

Smith, D. (2020, March 27). US surpasses China for highest number of confirmed Covid-19 cases in the world. *The Guardian.* Retrieved from https://www.theguardian.com/world/2020/mar/26/coronavirus-outbreak-us-latest-trump

WHO. (Forthcoming). *WHO Director-General’s opening remarks at the media briefing on COVID-19—March 11, 2020.* Retrieved June 21, 2020, from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-11-march-2020

---

**How to cite this article:** Tan T, Huang T, Wang X, Zuo Z. A preliminary investigation of COVID-19 transmission in the United States by incorporating social media sentiments. *Proc Assoc Inf Sci Technol.* 2020;57:e370. https://doi.org/10.1002/pra2.370