Opinion

Your midnight tweet could signal a new #pandemic

Da-Young(Diane) Kang
Undergraduate at Johns Hopkins University, USA

Correspondence: Da-Young(Diane) Kang, Undergraduate at Johns Hopkins University, USA, Tel 6672784252, Email dkang23@jhu.edu

"Day 3 of having a fever now… is this covid? #Covid19" When we type in and upload a few words that fit into 280 characters on Twitter, our posts become a part of an important database to detect infectious diseases. Such surveillance of the population is done by using machine learning to read through the twitter posts by keywords. Interestingly, statistical and machine learning techniques can catch signs of diseases beyond influenza-like illnesses (ILI), such as dengue, HIV, gastroenteritis, ebola, diarrhoea, and allergies. Based on the posts they read, they can measure the magnitude of the diseases or cases. Even more, machine learning techniques detect drug abuse, suicide attempts, and depressions. These methods in public health are celebrated for their uniqueness. However, there are doubts about relying on the methods which deviate from the mainstream data from the governments or medical organizations.[1]

Specifically for Twitter, there is a risk of misinformation because every post cannot be trusted as genuine. Also, the machine learning techniques have difficulties reading through nuances and sarcasms to collect accurate data.[2] Beyond Twitter, more concerns exist for companies who are experimenting with diverse methods too. First, there are unresolved problems in how to make the data available to the public. Although the National Institutes of Health require data to be shared publicly, the data is not easily shared due to ownership issues. Since sharing data is essential in reducing gaps among researchers,[3] the companies who are unwilling to publish their data may stagger the research processes. Another problem is that employing such data science methods may actually widen the health disparities in the populations. Since the methods heavily rely on internet users, the data may not represent the entire population accurately. Referred as "digital inequality," people from low socioeconomic status face challenges in using "computer technologies due to lack of relevant skills and resources"[4] or weak internet connection. Thus, "digital inequality" hinders gathering data that is representative of populations from all socioeconomic statuses.

Despite the limitations, various data science methods should continue to be studied and developed. This is because its data combined with the data from insurance companies, governmental surveys and studies, and medical institutions will allow researchers to have a better view of the prevalence of diseases. For example, a Canadian company BlueDot demonstrated the effectiveness of distinct surveillance methods at the dawn of the COVID-19 pandemic by obtaining the global airline ticketing data.[5] With the data, BlueDot recognized the signs of the pandemic even before the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO).[6] Also, BlueDot predicted the first eight cities which would encounter COVID-19 and published the first scientific paper on the novel coronavirus. The timely and accurate report was possible because the company took an exclusive approach to the available data: they examined where travelers from Wuhan most travel to based on the global airline ticketing data. This travel information helped them to predict where the coronavirus would spread next. BlueDot was successful in making predictions because it did not depend on local governmental officials at the outbreak site, who could be less eager to share information on infectious diseases.[6] Moreover, BlueDot

employed natural language processing and other machine learning tools to review news reports, blogs, and medical reports on animal diseases in 65 different languages. By utilizing artificial intelligence, BlueDot reported and predicted the magnitude of the coronavirus. Though they are still working to share their data to the public, their data will assist health officials around the world to take quick actions against the future pandemics.[6]

Another company in the U.S. called FluNearYou (FNY) is an online surveillance platform created in 2011 by American Public Health Association, HealthMap of Boston Children's Hospital, and the Skoll Global Threats Fund to predict and track influenza outbreaks early.[7] The company incorporates a crowdsourcing method, which in data science is gathering ideas or information from a large group of internet users rather than traditional employees.[8] Thus, FNY recruits volunteers from the U.S. and Canada to report ILI weekly by answering survey questions on what symptoms (i.e. fever, cough, and sore throat) they had in the previous week.[7,9] Research done by Baltrusaitis et al. shows that 65% of the FNY users do not seek medical attention regardless of their flu symptoms. This implies that FNY is taking counts of symptomatic people who are not caught by medical institutions. Such non-official numbers enable FNY to predict influenza transmission patterns more accurately than state and federal governments who turn to medical institutions for the data.[9] Other researchers supported that FNY is accurate in its timing and magnitude of the influenza outbreak when compared to the ILI rates from CDC ILI Surveillance Network.[7] Although FNY requires an adequate number of online participants for accurate data, their creative method can cover the missing numbers who do not visit medical institutions for their illnesses.[9]

Although the limitations of data sharing and digital inequality still exist in varying data science tools such as BlueDot and FNY, their limitations could be compensated by integrating their data to the official data from the governments and medical communities. Despite the existing doubts in the medical community, the "abnormal" data science methods should be accepted to augment the well-established public health surveillance methods to aid governmental officials to make quick, evidence-based decisions to protect the public from diseases. This is because the mindless tweet posted at midnight could be the first sign of the global pandemic.

Acknowledgments

None.

Conflicts of Interset

None.

References

1. Edo-Osagie O, De La Iglesia B, Lake I, et al. A scoping review of the use of twitter for public health research. *Comput Biol Med*. 2020;122:103770.

3. Pershad Y, Hangge PT, Albadawi H, et al. Social medicine: Twitter in healthcare. *J Clin Med*. 2018;7(6).

4. Zhang X, Pérez-Stable EJ, Bourne PE, et al. Big data science: Opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn Dis*. 2017;27(2):95-106.

5. McCloud RF, Okechukwu CA, Sorensen G, et al. Beyond access: Barriers to internet health information seeking among the urban poor. *J Am Med Inform Assoc.* 2016;23(6):1053-1059.

6. BlueDot: Outbreak risk software. BlueDot. 2021.

7. Niiler E. An AI epidemiologist sent the first alerts of the coronavirus. *Wired*. 2020.

8. Smolinski MS, Crawley AW, Baltrusaitis K, et al. Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons. *Am J Public Health*. 2015;105(10):2124-2130.

9. Definition of crowdsourcing. Merriam-Webster Web site. 2021.

10. Baltrusaitis K, Brownstein JS, Scarpino SV, et al. Comparison of crowd-sourced, electronic health records based, and traditional health-care based influenza-tracking systems at multiple spatial resolutions in the united states of america. *BMC Infect Dis*. 2018;18(403).