Letter to the Editor

Data mining can play a critical role in COVID-19 linked mental health studies

Dear editor,

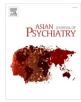At the time of writing (25th August 2020), the tally of global confirmed novel coronavirus (COVID-19) cases has exceeded 23.8 million (Worldometer, 2020). As the virus is still wreaking havoc in most parts of the world, this unprecedented global public health disaster also has had an enormous impact on people's mental health (Rajkumar, 2020; Tandon, 2020a; Vindegaard and Benros, 2020). While psychology researchers across the globe are adopting various research approaches and techniques in fighting against the pandemic, in this letter, I would like to introduce an emerging technique into COVID-19 linked mental health studies, viz., data mining.

Data mining is an interdisciplinary process which incorporates knowledge of computer science and statistics to analyze large observational datasets. The aim of data mining is to find unsuspected relationships or patterns from datasets and to summarize the data in novel ways (Hand et al., 2001). Data mining is a broad concept and encompasses a wide spectrum of analytical methods. This letter presents two common data mining-based techniques with empirical examples to prove their merits in assisting mental health research.

The first technique is topic modeling, which is a text-mining approach that extracts semantic information from a text database and discovers topics (themes) based on word co-occurrence analysis. A widely used topic modeling method is the Latent Dirichlet Allocation (LDA), an unsupervised algorithm which uses a three-layer hierarchical probabilistic model to identify latent topics (Blei et al., 2003). In an LDA model, a document (e.g., an article, an abstract or a paragraph) can be assigned to multiple topics with various proportions instead of assigning to just one topic. Drawing on the LDA technique, I analyzed 908 abstracts of COVID-19 related mental health and psychological research articles published by July 2020 and indexed in Scopus. I used the R package *ldatuning* and estimated a best fitting topic number (8) for the abstract corpus. Then I used the *topicmodels* package to identify the eight topics and calculate their respective topic proportions. The results suggest that health professionals' mental health during the pandemic was the most studied topic in the retrieved abstracts, the topic proportion is 17.13 %. Moreover, pandemic linked domestic and family violence and relationship abuse was the second most studied topic (13.32 %). In addition to analysis of research articles, future studies could extend to analyze messages from social networking sites. For example, my colleagues and I are planning to use the LDA technique to investigate the Twitter posts of some verified psychologists and psychiatrists' accounts which were posed during the lockdown period. A limitation of LDA is that although the analysis is algorithm-based, it still requires researchers to manually summarize and label the topics. Which will inevitably contain subjective biases.

Another data mining technique is the analysis of Internet search behaviors. About 80 percent of Internet users have searched for online health information (Grohol et al., 2014). During the pandemic lockdown and under the strict social distancing restrictions, people are more reliant on the Internet than ever before and therefore are likely to instigate increased Internet search if they undergo mental health issues. As the world's largest search engine (Gupta et al., 2017), Google's search data can provide abundant information to predict and evaluate epidemics such as COVID-19 (Ayyoubzadeh et al., 2020). Using the *Google Trends* service, I conducted a series of correlation analyses of the relationships between COVID-19 prevalence or case growth and people's Google search behavior. For example, I found a significant relationship between COVID-19 case growth (in early July 2020) and people's recent Google search interests for the keyword "coronavirus" across the fifty states and the capital of the United States ($r = 0.66$, $p < 0.01$). Also, using the "related topics" and "related queries" features on the Google Trends website, I analyzed the relationships between an array of negative emotion keywords and regional COVID-19 case growth (in early August) in the US states. Among the twelve negative emotion words that I tested, ten were found to be positively related to case growth at the 0.1 significance level (83.3 %). For instance, "gloomy" ($r = 0.53$, $p < 0.01$), "cry" ($r = 0.45$, $p < 0.01$) and "depressed" ($r = 0.43$, $p < 0.01$) were the top three search words. However, when I attempted to test some more aggressive search words such as "violent" and "suicide", significant relationships were not found. The results suggest that although people are experiencing low mood and increased stress due to the pandemic, most people are unlikely to take extreme actions under the current situation.

Data mining can exert greater influence on COVID-19 linked mental health studies if we continue to underline its importance. For example, sentiment analysis can be implemented to evaluate Internet users' overall sentiment status and thereby to monitor public health concerns (Singh et al., 2020). Novel data mining techniques like the Correlation Explanation learning algorithm (Li et al., 2020) can be used in modeling spatiotemporal patterns of mental disorder symptoms. Lastly, but most importantly, due to the varying quality of COVID-19 related data collected across countries, we must be exceptionally cautious in mining and interpreting data (Tandon, 2020b).

Financial disclosure

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgement

None.

References

Ayyoubzadeh, S., Ayyoubzadeh, S., Zahedi, H., Ahmadi, M., Kalhori, S., 2020. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill. 6 (2), 192–198. https://doi.org/10.2196/18828.

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (Jan), 993–1022.

Grohol, J., Slimowicz, J., Granda, R., 2014. The quality of mental health information commonly searched for on the internet. Cyberpsych. Beh. Soc. N. 17 (4), 216–221.

Gupta, H., Singh, S., Sinha, P., 2017. Multimedia tool as a predictor for social media advertising – a YouTube way. Multimed. Tools Appl. 76 (18), 18557–18568.

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. The MIT Press, Cambridge, US.

Li, D., Chaudhary, H., Zhang, Z., 2020. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. Int. J. Environ. Res. Public Health 17 (14), 4988. https://doi.org/10.3390/ijerph17144988.

Rajkumar, R.P., 2020. COVID-19 and mental health: a review of the existing literature. Asian J. Psychiatr. 52, 102066.

Singh, P., Singh, S., Sohal, M., Dwivedi, Y., Kahlon, K., Sawhney, R., 2020. Psychological fear and anxiety caused by COVID-19: insights from twitter analytics. Asian J. Psychiatr. 54, 102280.

Tandon, R., 2020a. The COVID-19 pandemic, personal reflections on editorial responsibility. Asian J. Psychiatr. 50, 102100.

Tandon, R., 2020b. COVID-19 and mental health: preserving humanity, maintaining sanity, and promoting health. Asian J. Psychiatr. 51, 102256.

Vindegaard, N., Benros, M., 2020. COVID-19 pandemic and mental health consequences: systematic review of the current evidence. Brain Behav. Immu. https://doi.org/10.1016/j.bbi.2020.05.048. In press.

Worldometer, 2020. Countries Where COVID-19 has Spread. Available from www.worldometer.info, (Accessed 25 August 2020).

Tenghao Zhang

School of Business and Law, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Australia
E-mail address: tenghaoz@our.ecu.edu.au.