Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210342

Integrated Infodemic Surveillance System: The Case of COVID-19 in South Korea

Gil-sung PARK[a], Jintae BAE[a], Jong Hun LEE[a], Byung Yeon YUN[a], Byunghwee LEE[b] and Eun Kyong SHIN[a,1]
[a] *Department* of Sociology, Korea University, South Korea
[b] *Department of Physics, Korea Advanced Institute of Science and Technology, South Korea*

Abstract. This study merges multiple COVID-19 data sources from news articles and social media to propose an integrated infodemic surveillance system (IISS) that implements infodemiology for a well-tailored epidemic management policy. IISS is an à-la-carte infodemic surveillance solution that enables users to gauge the epidemic related consensus, which compiles epidemic-related data from multiple sources and equipped with various methodological toolkits - topic modeling, Word2Vec, and social network analysis. IISS can provide reliable empirical evidence for proper policymaking. We demonstrate the heuristic utilities of IISS using empirical data from the first wave of COVID-19 in South Korea. Measuring discourse congruence allows us to gauge the distance between the discourse corpus from different sources, which can highlight consensus and conflicts in epidemic discourse. Furthermore, IISS detects discrepancies between social concerns and main actors.

Keywords. COVID-19, Infodemic, Pandemic, Applied Network Science, Health Surveillance, Population Health, Health Informatics

1. Introduction

The growing complexity and fatality of coronavirus disease 2019 (COVID-19) [1-3] urges the development of new ways of understanding and dealing with the pandemic. This global pandemic is not an unprecedented phenomenon. However, the spread of unknown viruses, both with a high reproductive number (R0) and a high case fatality rate (CFR) in the digitized world presents a novel challenge [4, 5]. While the majority of people experience the pandemic without direct contact with the virus, everyone is exposed to information about the virus from a variety of sources; hence, information structures are critical for determining the quality of our daily pandemic experience [6]. Since the emergence of severe acute respiratory syndrome (SARS) in 2003, the 'infodemic age' has become an inevitable battle field [7, 8] and the magnitude of infodemics has expanded exponentially [9].

The COVID-19 pandemic is beyond a public 'health' emergency of international concern. Infectious disease is not a mere pathological matter but a social disaster. The longer the pandemic persists, the more complex the infodemic becomes, and the higher its social costs. An overabundance of information with high uncertainty brews high

[1] Eun Kyong Shin; Korea University, Department of Sociology, Seoul, South Korea E-mail: eunshin@korea.ac.kr.

levels of social distrust that can impede the efficient deliberation where it's most needed. The unfolding of COVID-19 has engendered a global social disaster off-line and on-line that requires an inclusive and integrated infodemic surveillance strategy to cope with the pandemic with minimum social disturbance [10]. However, existing studies of infodemics have focused on segmental comparison rather than actively grafting multiple information sources.

In this study, merging multiple COVID-19 data sources from news articles and social media, which are the prime areas of concern for WHO, we demonstrate the heuristic utilities of the integrated infodemic surveillance system (IISS), implementing infodemiology [11-13] to a well-tailored and consensus-focused epidemic management policy. IISS is an à-la-carte infodemic surveillance solution that enables users to gauge the epidemic related consensus, which compiles epidemic-related data from multiple sources and equipped with various methodological toolkits. IISS is an infrastructure instrument used to depict and examine pandemic-related discourse.

2. Data and Methods

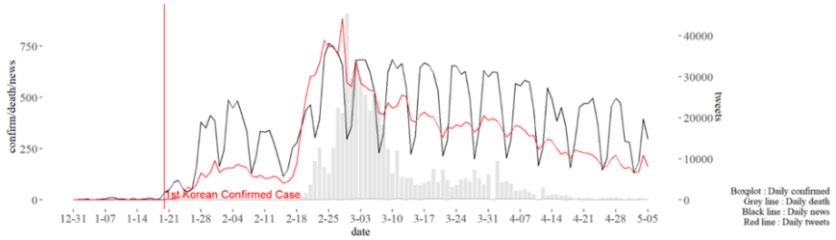

Figure 1. Descriptive Trajectory of Data

This study empirically examines the social orchestration of COVID-19 in South Korea (hereafter, K-COVID-19) during the first wave (from January 1st to May 15th, 2020) of the virus. Two different types of data were compiled for this study. First, we collected news articles from five major media outlets using a keyword-based search. The K-COVID-19 news text corpora allowed us to identify the main concerns related to COVID-19. The dataset is composed of 43,057 news articles. In addition, the same keyword-based search of Twitter data led to a total of 1,724,027 tweets composed of 1,295,151 (75.55%) original tweets and 419,108 (24.45%) retweets. For each Twitter entry, information related to the poster's total number of followers and following, the date of entry of the tweet, retweets, and the inclusion of external media were collected. We use the integrated text corpora to investigate the K-COVID-19 discourse. The basic descriptive statistics are shown in Figure 1.

To examine the K-COVID-19 corpora, we combined three different methods in IISS. First, we implemented topic modeling [14, 15], and traced how the main topics around K-COVID-19 evolved over time. Next, using Word2Vec [16], we explore word embedding to identify the keywords that were most closely related to a given focal concept, such as quarantine policy or facemasks. By identifying closely related neighbor words, we reconstructed the main concern networks and compared Word2Vec networks to gauge congruence sentiments in different data sources. Nodes in the network are the key concepts (top thirty neighbor concepts in the cosine distance), and the edge represents the similarity between the two words. The smaller the congruence, the wider

the gap between the news media and social media, which can result in challenges for achieving efficient social consensus. Measuring discourse congruence can allow us to gauge the distance between the discourse corpus from different sources, which can highlight the consensus and conflicts in epidemic discourse. Lastly, we construct main agent networks based on social network analysis [17]. Extracting individuals mentioned in news articles, social network analysis can identify the central actors in the handling of this infectious disease. Nodes in the networks are the actors, and the line between the nodes is measured by their co-appearance in an article. A total of 3,995 individuals were mentioned in the articles; 1,387 individuals from the political sector, 700 individuals from the economic sector, 1,478 individuals from the civic sector, 364 individuals from the medical sector, and 66 individuals from the government quarantine authorities. There were 31,390 coappearance ties. Data collection and analyses were conducted using Python (V 3.7), R (V 4.0.3), and Gephi (0.9.2).

3. Results

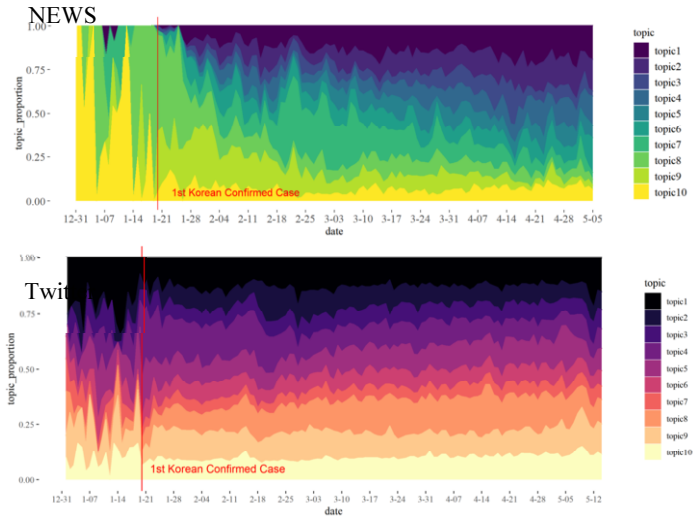

Figure 2. Topic Modeling Trajectories in News and Twitter

The results of topic modeling show the different sensitivities of the topic in two different sources. The news media topic content trajectories (Figure 2 Top) were different from the Twitter topic content trajectories (Figure 2 Bottom). News topic dominance has moved from the medical topic (Topic 10 in Figure 2 Top) to the political one (Topic 1 in Figure 2 Top) and Twitter topic focus has shifted from social encouragement (Topic 1 in Figure 2 Bottom) to social distancing (Topic 10 in in Figure 2 Bottom). IISS topic modeling help us identifying the social focus zone and the convergence and differences between different discourse sources.

Congruence measures from the Word2Vec results are shown in Figure 3. Orange color nodes denote the overlapping neighbor topics. The congruency for COVID-19 (0.33, Figure 3 Left) is smaller than the congruency for masks (0.43, Figure 3 Right), indicating that public opinion converges more regarding facial masks. Neighboring concept networks for Twitter data (COVID-19 0.379; Mask 0.459) are more densely connected than News data (COVID-19 0.357; Mask 0.397). IISS congruency detection

can be used to detect the contour of social consensus and to identify which topic is most contested and which topic is least disputed.

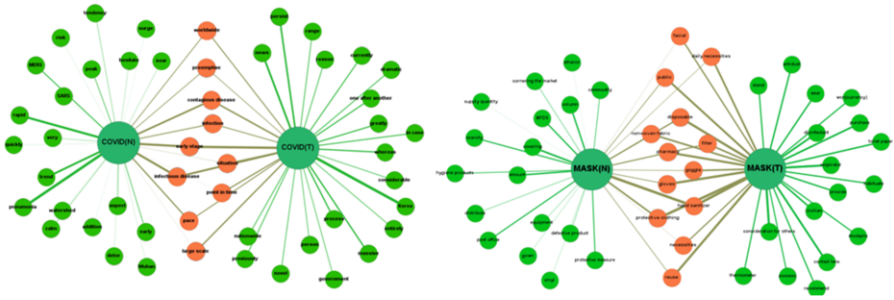

Figure 3. Topic Congruency Network between News and Twitter

The results of the major agent network highlight the discrepancies between social concerns and main actors. The results of topic modeling show that the central concerns related to COVID-19 were densely populated with social and economic issues. However, the actors that appeared in the COVID-19 agency networks were mainly government officials and politicians. Throughout the 1st wave on K-COVID-19, political sector centralities were outstandingly higher than agents from other fields. Overall, political sector's weighted centrality was 53.46, whereas medical sector's weighted centrality was 6.19. Voices from social scientists were rare, if not entirely missing. While topics related to COVID-19 cover a wide range of social issues, from its nascent stage (Figure 4 Left, from Jan 1st to Feb 17th 2020), the most dominant K-COVID-19 agents were from the government: experts in the social science domain rarely appeared. At the early stage (Figure 4 Right, from Feb 18th to May 5th), the most central actors were politicians. As the pandemic had lasted longer than expected, and medical experts emerged linking politicians in different sectors. Moreover, we found the presence of strong homophily among the K-COVID-19 agents: 75.63% of the ties were homogeneous (links within the same sector) and 21.27% of the ties were heterogeneous (links between different sectors).

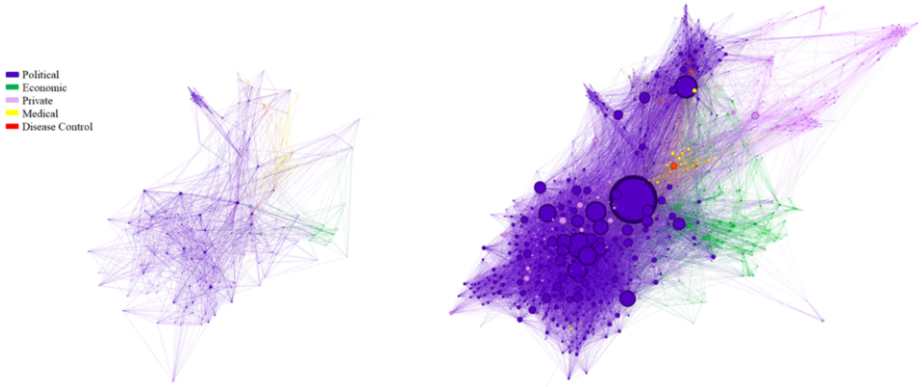

Figure 4. Major Agents Network Evolution

4. Conclusion

This study articulates the heuristic benefits of implementing integrated infodemiology for epidemic health surveillance. Infodemic can be amplified through digital platforms and spread further and faster than the virus and the infectious disease control in the digital era, infodemic surveillance is a pivotal part of the pandemic control. Well-designed data curation and analysis is critical for efficient health resource circulation and distribution [18, 19]. We designed the IISS to heuristically use the digital COVID-19 corpus to collect balanced data and provide reliable empirical evidence for proper policymaking. A pandemic is a social catastrophe, and social scientists are the critical pieces of the puzzle. Thus, adequate and appropriate pandemic risk management, particularly at a global scale, should involve a larger and inclusive expert community and secure diverse communication channels.

References

[1] Paules CI, Marston HD, Fauci AS. Coronavirus infections—more than just the common cold. Jama. 2020;323(8):707-708.
[2] Fauci AS, Lane HC, Redfield RR. Covid-19—navigating the uncharted, ed: Mass Medical Soc, 2020.
[3] W. H. Organization. Coronavirus disease (COVID-19). 2020.
[4] Zarocostas J. How to fight an infodemic. The Lancet. 2020;395(10225):676.
[5] Eysenbach G. How to fight an infodemic: the four pillars of infodemic management. Journal of medical Internet research. 2020;22(6):e21820.
[6] Lee B, Jeong H, Shin EK. On-line (TweetNet) and Off-line (EpiNet): The Distinctive Structures of the Infectious. Spriger Studies in Computational Intelligence. 2020.
[7] Coxe D. The new infodemic age. Maclean's. 2003 June;9:36.
[8] Rothkopf DJ, When the buzz bites back, The Washington Post.2003;11:B1&B5.
[9] Gazendam A, Ekhtiari S, Wong E, Madden K, et al, The "Infodemic" of journal publication associated with the novel coronavirus disease. JBJS. 2020;102(13):e64, 2020.
[10] Shin EK, Lee JH. Evidence-Based Health Intelligence with Globally Localized Epidemic Knowledge Base: Merging Pathological Data, Socio-Environmental Data and Intervention Knowledge Data. Stud Health Technol Inform. 2020 Jun 26;272:17-20.
[11] Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009 Mar 27;11(1):e11.
[12] Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc. 2006;2006. p.244-8.
[13] Eysenbach G. Infodemiology: The epidemiology of (mis)information. Am J Med. 2002 Dec 15;113(9):763-5.
[14] Hong L, Davison BD, Empirical study of topic modeling in twitter, in Proceedings of the first workshop on social media analytics; 2010. p. 80-88.
[15] Wallach HM, Topic modeling: beyond bag-of-words, in Proceedings of the 23rd international conference on Machine learning;2006. p. 977-984.
[16] Rong X. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738,* 2014.
[17] Scott J. Social network analysis. Sage. 2017.
[18] Shaban-Nejad A, Kamaleswaran R, Shin EK, Akbilgic O. Health intelligence. in *B*iomedical Information Technology: Elsevier; 2020. p. 197-215.
[19] Shin EK, Shaban-Nejad A. Public Health Intelligence and the Internet: Current State of the Art. in Public Health Intelligence and the Internet, A. Shaban-Nejad, J. S. Brownstein, and D. L. Buckeridge Eds. Cham: Springer International Publishing; 2017. p. 1-17.