The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0737-8831.htm

Retrospective and prospective approaches of coronavirus publications in the last half-century: a Latent Dirichlet allocation analysis

Received 2 September 2020
Revised 13 December 2020
11 March 2021
Accepted 13 March 2021

Farshid Danesh
Information Management Research Department, Regional Information Center for Science and Technology, Shiraz, Iran, and
Meisam Dastani and Mohammad Ghorbani
Gonabad University of Medical Sciences, Gonabad, Iran

Abstract

Purpose – The present article's primary purpose is the topic modeling of the global coronavirus publications in the last 50 years.

Design/methodology/approach – The present study is applied research that has been conducted using text mining. The statistical population is the coronavirus publications that have been collected from the Web of Science Core Collection (1970–2020). The main keywords were extracted from the Medical Subject Heading browser to design the search strategy. Latent Dirichlet allocation and Python programming language were applied to analyze the data and implement the text mining algorithms of topic modeling.

Findings – The findings indicated that the SARS, science, protein, MERS, veterinary, cell, human, RNA, medicine and virology are the most important keywords in the global coronavirus publications. Also, eight important topics were identified in the global coronavirus publications by implementing the topic modeling algorithm. The highest number of publications were respectively on the following topics: "structure and proteomics," "Cell signaling and immune response," "clinical presentation and detection," "Gene sequence and genomics," "Diagnosis tests," "vaccine and immune response and outbreak," "Epidemiology and Transmission" and "gastrointestinal tissue."

Originality/value – The originality of this article can be considered in three ways. First, text mining and Latent Dirichlet allocation were applied to analyzing coronavirus literature for the first time. Second, coronavirus is mentioned as a hot topic of research. Finally, in addition to the retrospective approaches to 50 years of data collection and analysis, the results can be exploited with prospective approaches to strategic planning and macro-policymaking.

Keywords Coronavirus, COVID-19, Text-mining, Topic modeling, Scientometrics, Web of science core collection (WOSCC), Latent Dirichlet allocation (LDA)

Paper type Research paper

Researchers appreciate the Vice-Chancellor for Research and Technology of Gonabad University of Medical Sciences for its financial support and the respiratory infectious diseases specialists for their expert advice.

Research funding: The present article is extracted from a research project with code A-10–1263–2 and research ethics ID. I.R.GMU.REC.1398.189 approved by the Research Council of Gonabad University of Medical Sciences and conducted with financial support.

Library Hi Tech
© Emerald Publishing Limited
0737-8831
DOI 10.1108/LHT-09-2020-0216

References

Abramson, D., Lees, M., Krzhizhanovskaya, V.V., Dongarra, J.J. and Sloot, P.M. (2014), "Big data meets computational science, preface for ICCS 2014", *ICCS*, pp. 1-7.

Abuhay, T.M., Kovalchuk, S.V., Bochenina, K.O., Kampis, G., Krzhizhanovskaya, V.V. and Lees, M.H. (2017), "Analysis of computational science papers from ICCS 2001-2016 using topic modeling and graph theory", *Procedia Computer Science*, Vol. 108, pp. 7-17.

Amawi, H., Abu Deiab, G.A.I., A Aljabali, A.A., Dua, K. and Tambuwala, M.M. (2020), "COVID-19 pandemic: an overview of epidemiology, pathogenesis, diagnostics and potential vaccines and therapeutics", *Therapeutic Delivery*, Vol. 11 No. 4, pp. 245-268.

Ari, N. and Ustazhanov, M. (2014), "Matplotlib in python", *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)*, IEEE, pp. 1-6.

Berry, M.W. and Kogan, J. (2010), *Text Mining. Applications and Theory*, John Wiley & Sons, West Sussex.

Bertram, S., Heurich, A., Lavender, H., Gierer, S., Danisch, S., Perin, P., Lucas, J.M., Nelson, P.S., Pöhlmann, S. and Soilleux, E.J. (2012), "Influenza and SARS-coronavirus activating proteases TMPRSS2 and HAT are expressed at multiple sites in human respiratory and gastrointestinal tracts", *PloS One*, Vol. 7 No. 4, p. e35876.

Birkle, C., Pendlebury, D.A., Schnell, J. and Adams, J. (2020), "Web of Science as a data source for research on scientific and scholarly activity", *Quantitative Science Studies*, Vol. 1 No. 1, pp. 363-376.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent Dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, January, pp. 993-1022.

Blei, D.M. (2012), "Probabilistic topic models", *Communications of the ACM*, Vol. 55 No. 4, pp. 77-84.

Bonilla-Aldana, D.K., Quintero-Rada, K., Montoya-Posada, J.P., Ramírez-Ocampo, S., Paniz-Mondolfi, A., Rabaan, A.A., Sah, R. and Rodríguez-Morales, A.J. (2020), "SARS-CoV, MERS-CoV and now the 2019-novel CoV: have we investigated enough about coronaviruses?–A bibliometric analysis", *Travel Medicine and Infectious Disease*, Vol. 33, 101566.

Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N. and Börner, K. (2011), "Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches", *PloS One*, Vol. 6 No. 3, p. e18029.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. and Blei, D.M. (2009), "Reading tea leaves: how humans interpret topic models", in *Neural Information Processing Systems*, Vol. 22, pp. 288-296.

Chen, K., Kou, G., Shang, J. and Chen, Y. (2015), "Visualizing market structure through online product reviews: integrate topic modeling, TOPSIS, and multi-dimensional scaling approaches", *Electronic Commerce Research and Applications*, Vol. 14 No. 1, pp. 58-74.

Cheng, X., Shuai, C., Liu, J., Wang, J., Liu, Y., Li, W. and Shuai, J. (2018), "Topic modeling of ecology, environment and poverty nexus: an integrated framework", *Agriculture, Ecosystems & Environment*, Vol. 267, pp. 1-14.

Colavizza, G., Costas, R., Traag, V.A., Van Eck, N.J., Van Leeuwen, T. and Waltman, L. (2021), "A scientometric overview of CORD-19", *PloS One*, Vol. 16 No. 1, p. e0244839.

Dancy-Scott, N., Dutcher, G.A., Keselman, A., Hochstein, C., Copty, C., Ben-Senia, D., Rajan, S., Asencio, M.G. and Choi, J.J. (2018), "Trends in HIV terminology: text mining and data visualization assessment of International AIDS conference abstracts over 25 years", *JMIR public health and surveillance*, Vol. 4 No. 2, p. e50.

Danesh, F. and Ghavidel, S. (2020), "Coronavirus: scientometrics of 50 Years of global scientific productions", *Iranian Journal of Medical Microbiology*, Vol. 14 No. 1, pp. 1-16.

Danesh, F., Ghavidel, S. and Piranfar, V. (2020), "Coronavirus: discover the structure of global knowledge, hidden patterns & emerging events", *Journal of Advances in Medical and Biomedical Research*, Vol. 28 No. 130, pp. 253-264.

Dehghanbanadaki, H., Seif, F., Vahidi, Y., Razi, F., Hashemi, E., Khoshmirsafa, M. and Aazami, H. (2020), "Bibliometric analysis of global scientific research on Coronavirus (COVID-19)", *Medical Journal of the Islamic Republic of Iran*, Vol. 34, p. 51.

Frakes, W.B. and Baeza-Yates, R. (1992), *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, Saddle River, NJ.

Greene, D. and Cross, J.P. (2017), "Exploring the political agenda of the European parliament using a dynamic topic modeling approach", *Political Analysis*, Vol. 25 No. 1, pp. 77-94.

Greene, D., O'Callaghan, D. and Cunningham, P. (2014), "How many topics? Stability analysis for topic models", *Joint European conference on machine learning and knowledge discovery in databases*, Springer, Berlin, Heidelberg, pp. 498-513.

Griffiths, T.L. and Steyvers, M. (2004), "Finding scientific topics", *Proceedings of the National Academy of Sciences*, Vol. 101, suppl 1, pp. 5228-5235.

Hafeez, A., Ahmad, S., Siddqui, S.A., Ahmad, M. and Mishra, S. (2020), "A review of COVID-19 (Coronavirus Disease-2019) diagnosis, treatments and prevention", *EJMO*, Vol. 4 No. 2, pp. 116-125.

Haghani, M., Bliemer, M.C., Goerlandt, F. and Li, J. (2020), "The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: a scientometric analysis and scoping review", *Safety Science*, Vol. 129, 104806.

Hashimi, H., Hafez, A. and Mathkour, H. (2015), "Selection criteria for text mining approaches", *Computers in Human Behavior*, Vol. 51, pp. 729-733.

Hossain, M.M. (2020), *Current Status of Global Research on Novel Coronavirus Disease (Covid-19): A Bibliometric Analysis and Knowledge Mapping. Hossain MM. Current Status of Global Research on Novel Coronavirus Disease (COVID-19): A Bibliometric Analysis and Knowledge Mapping [version 1]*.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L. (2019), "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey", *Multimedia Tools and Applications*, Vol. 78 No. 11, pp. 15169-15211.

Jensen, P.B., Jensen, L.J. and Brunak, S. (2012), "Mining electronic health records: towards better research applications and clinical care", *Nature Reviews Genetics*, Vol. 13 No. 6, pp. 395-405.

Kagan, D., Moran-Gilad, J. and Fire, M. (2020), "Scientometric trends for coronaviruses and other emerging viral infections", *GigaScience*, Vol. 9 No. 8, p. giaa085.

Katre, P.D. (2019), "Text mining and comparative visual analytics on large collection of speeches to trace socio-political issues", *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, IEEE, pp. 108-114.

Kim, Y.M. and Delen, D. (2018), "Medical informatics research trend analysis: a text mining approach", *Health Informatics Journal*, Vol. 24 No. 4, pp. 432-452.

Ksiazek, T.G., Erdman, D., Goldsmith, C.S., Zaki, S.R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J.A., Lim, W. and Rollin, P.E. (2003), "A novel coronavirus associated with severe acute respiratory syndrome", *New England Journal of Medicine*, Vol. 348 No. 20, pp. 1953-1966.

Lam, C., Lai, F.C., Wang, C.H., Lai, M.H., Hsu, N. and Chung, M.H. (2016), "Text mining of journal articles for sleep disorder terminologies", *PloS One*, Vol. 11 No. 5, p. e0156031.

Lamba, M. and Madhusudhan, M. (2019), "Mapping of topics in DESIDOC journal of library and information Technology, India: a study", *Scientometrics*, Vol. 120 No. 2, pp. 477-505.

Li, J. and Zhang, K. (2007), "Keyword extraction based on TF-IDF for Chinese news document", *Wuhan University Journal of Natural Sciences*, Vol. 12 No. 5, pp. 917-921.

Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016), "An overview of topic modeling and its current applications in bioinformatics", *SpringerPlus*, Vol. 5 No. 1, pp. 1-22.

Manning, C.D., Raghavan, P. and Schütze, H. (2008), "XML retrieval", *An Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

MeSH Browser (2021), Available at: https://meshb.nlm.nih.gov/search (accessed 7 March 2021).

O'callaghan, D., Greene, D., Carthy, J. and Cunningham, P. (2015), "An analysis of the coherence of descriptors in topic modeling", *Expert Systems with Applications*, Vol. 42 No. 13, pp. 5645-5657.

Ozaydin, B., Zengul, F., Oner, N. and Delen, D. (2017), "Text-mining analysis of mHealth research", *mHealth*, Vol. 3 No. 12.

Park, A. and Conway, M. (2017), "Tracking health-related discussions on Reddit for public health applications", *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, Vol. 2017, p. 1362.

Peiris, J.S.M., Lai, S.T., Poon, L.L.M., Guan, Y., Yam, L.Y.C., Lim, W., Nicholls, J., Yee, W.K.S., Yan, W.W., Cheung, M.T. and Cheng, V.C.C. (2003), "Coronavirus as a possible cause of severe acute respiratory syndrome", *The Lancet*, Vol. 361 No. 9366, pp. 1319-1325.

Poutanen, S.M., Low, D.E., Henry, B., Finkelstein, S., Rose, D., Green, K., Tellier, R., Draker, R., Adachi, D., Ayers, M. and Chan, A.K. (2003), "Identification of severe acute respiratory syndrome in Canada", *New England Journal of Medicine*, Vol. 348 No. 20, pp. 1995-2005.

Radanliev, P., De Roure, D. and Walton, R. (2020), "Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development-In the first wave of the Covid-19 pandemic", *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, Vol. 14 No. 5, pp. 1121-1132.

Ram, S. (2020), "Coronavirus research trends: a 50–year bibliometric assessment", *Science and Technology Libraries*, Vol. 39 No. 2, pp. 210-226.

Rebholz-Schuhmann, D., Oellrich, A. and Hoehndorf, R. (2012), "Text-mining solutions for biomedical research: enabling integrative biology", *Nature Reviews Genetics*, Vol. 13 No. 12, pp. 829-839.

Rehurek, R. and Sojka, P., (2010), "Software framework for topic modeling with large corpora", *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*.

Rehurek, R. and Sojka, P. (2011), "Gensim—statistical semantics in python. Statistical semantics; Gensim", *Python; LDA; SVD*.

Röder, M., Both, A. and Hinneburg, A. (2015), "Exploring the space of topic coherence measures", *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399-408.

Rodriguez-Esteban, R. and Bundschus, M. (2016), "Text mining patents for biomedical knowledge", *Drug Discovery Today*, Vol. 21 No. 6, pp. 997-1002.

Rogers, J.R., Mills, H., Grossman, L.V., Goldstein, A. and Weng, C. (2020), "Understanding the nature and scope of clinical research commentaries in PubMed", *Journal of the American Medical Informatics Association*, Vol. 27 No. 3, pp. 449-456.

Rusanov, A., Miotto, R. and Weng, C. (2018), "Trends in anesthesiology research: a machine learning approach to theme discovery and summarization", *JAMIA Open*, Vol. 1 No. 2, pp. 283-293.

Sa'ed, H.Z. (2016), "Global research trends of Middle East respiratory syndrome coronavirus: a bibliometric analysis", *BMC Infectious Diseases*, Vol. 16 No. 1, pp. 1-7.

Saheb, T. and Saheb, M. (2019), "Analyzing and visualizing knowledge structures of health informatics from 1974 to 2018: a bibliometric and social network analysis", *Healthcare informatics research*, Vol. 25 No. 2, pp. 61-72.

Salloum, S.A., Al-Emran, M., Monem, A.A. and Shaalan, K. (2018), "Using text mining techniques for extracting information from research articles", in *Intelligent Natural Language Processing: Trends and Applications*, Springer, Cham, pp. 373-397.

Sbalchiero, S. and Eder, M. (2020), "Topic modeling, long texts and the best number of topics. Some Problems and solutions", *Quality and Quantity*, pp. 1-14.

Sharma, D., Kumar, B. and Chand, S. (2018), "Trend analysis in machine learning research using text mining", *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, IEEE, pp. 136-141.

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., Zhu, H., Zhao, W., Han, Y. and Qin, C. (2019), "From SARS to MERS, thrusting coronaviruses into the spotlight", *Viruses*, Vol. 11 No. 1, p. 59.

Sperandeo, R., Messina, G., Iennaco, D., Sessa, F., Russo, V., Polito, R., Monda, V., Monda, M., Messina, A., Mosca, L.L. and Mosca, L. (2020), "What does personality mean in the context of mental

health? A topic modeling approach based on abstracts published in PubMed over the last 5 years", *Frontiers in Psychiatry*, Vol. 10, p. 938.

Sullivan, H.C. and Roback, J.D. (2020), "Convalescent plasma: therapeutic hope or hopeless strategy in the SARS-CoV-2 pandemic", *Transfusion Medicine Reviews*, Vol. 34 No. 3, pp. 145-150.

Syed, S. and Weber, C.T. (2018), "Using machine learning to uncover latent research topics in fishery models", *Reviews in Fisheries Science and Aquaculture*, Vol. 26 No. 3, pp. 319-336.

Tran, B.X., Nguyen, L.H., Turner, H.C., Nghiem, S., Vu, G.T., Nguyen, C.T., Latkin, C.A., Ho, C.S. and Ho, R.C. (2019), "Economic evaluation studies in the field of HIV/AIDS: bibliometric analysis on research development and scopes (GAP RESEARCH)", *BMC Health Services Research*, Vol. 19 No. 1, p. 834.

Tran, B.X., Phan, H.T., Nguyen, Q.N., Ewedairo, O.A., Vu, G.T., Ha, G.H., Nguyen, T.P., Latkin, C.A., Ho, C.S. and Ho, R.C. (2020), "Pre-exposure prophylaxis in HIV research: a latent Dirichlet allocation analysis (GAPRESEARCH)", *AIDS Reviews*, Vol. 22 No. 2, pp. 1-9.

Vorontsov, K., Frei, O., Apishev, M., Romov, P. and Dudarenko, M. (2015), "Bigartm: open source library for regularized multimodal topic modeling of large collections", *International Conference on Analysis of Images, Social Networks and Texts*, Springer, Cham, pp. 370-381.

Wang, C., Blei, D. and Heckerman, D. (2008), "Continuous-time dynamic topic models", *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 579-586.

Wieczorek, O. and Schubert, D. (2020), *The Symbolic Power of the Research Excellence Framework. Evidence from a Case Study on the Individual and Collective Adaptation of British Sociologists*, Bamberg University Press, Bamberg.

Xiao, F., Tang, M., Zheng, X., Liu, Y., Li, X. and Shan, H. (2020), "Evidence for gastrointestinal infection of SARS-CoV-2", *Gastroenterology*, Vol. 158 No. 6, pp. 1831-1833.

Yau, C.K., Porter, A., Newman, N. and Suominen, A. (2014), "Clustering scientific documents with topic modeling", *Scientometrics*, Vol. 100 No. 3, pp. 767-786.

Zaki, A.M., Van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D. and Fouchier, R.A. (2012), "Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia", *New England Journal of Medicine*, Vol. 367 No. 19, pp. 1814-1820.

Zhang, Y., Chen, M. and Liu, L. (2015), "A review on text mining", *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, pp. 681-685.

Zhao, J., Yuan, Q., Wang, H., Liu, W., Liao, X., Su, Y., Wang, X., Yuan, J., Li, T., Li, J. and Qian, S. (2020), "Antibody responses to SARS-CoV-2 in patients with novel coronavirus disease 2019", *Clinical Infectious Diseases*, Vol. 71 No. 16, pp. 2027-2034.

About the authors

Farshid Danesh has a BA, MS and Ph.D. in Knowledge and Information Sciences. He obtained his Ph.D. degree from the Ferdowsi University of Mashhad in 2013. He has been teaching and researching in various fields of scientometrics, text mining and scholarly communication since 2006. He is now an Assistant Professor in Information Management Research Department at Regional Information Center for Science and Technology (RICeST), Shiraz, Iran.

Meisam Dastani is currently Ph.D. in Knowledge and Information Science (KIS) in Iran. He is the director of Information Services and Knowledge Translation at Gonabad University of Medical Sciences. Data and text mining, Natural language process (NLP), and web evaluation technics and Software are his research interests. Meisam Dastani is the corresponding author and can be contacted at: meisam.dastani@gmail.com

Mohammad Ghorbani is a Ph.D. candidate in hematology and blood transfusion in Iran. He is a faculty member in the Pathology Department, medical school, Gonabad University of Medical Sciences.

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**