

Natural Language Processing-Based Solution for the Management of Covid-19 Infodemic

G.A.D. Ganepola

*Department of Information Technology, Faculty of Computing,
General Sir John Kotelawala Defense University, Ratmalana 10390, Sri Lanka
dasunig@kdu.ac.lk*

The Covid-19 pandemic outbreak occurred when technology is expeditious for information generation and sharing. This created an “Infodemic”; a continuous amplification of health information leading to an information overload. This has challenged the human population because the mismanagement of information disrupts the processes of preventing the virus from the spread and safeguarding the public resulting in decreasing effectiveness to end this pandemic situation. A key segment of the Infodemic is medical and scientific research publications related to the virus and the pandemic. Research has been conducted countlessly worldwide since the outbreak leading to immeasurable numbers of publications, which has made scientists difficult to keep pace with ongoing and potential research related to Covid-19. Hence, urgent assistance for information management is required. This research aims at information management through Artificial Intelligence utilizing the technologies in Natural Language Processing. The objective of this research is to develop a model that can discover abstract topics and themes in the English language with respect to Covid-19 related scientific publications by text analytics. Covid-19 Open Research Dataset was used for this research. Publications during the period from January 2020 to January 2021 were selected for the discovery of contemporary topics and themes. The abstracts of those publications were selected as they are summaries of publications. Thereafter, abstracts were tested for text similarity by Cosine similarity metric and were grouped based on the score. Afterward, data cleaning process, removing punctuation, lower casing, and stop words were carried out. Tokenization & Lemmatization and creating trigrams were performed before generating the Corpus & Dictionary. The topic model was developed from Latent Dirichlet Algorithm, and hyper-parameter tuning was performed for model optimization. Model performance was evaluated through coherence score and the initial score resulted in 0.6501. Continuous model performance evaluation is being performed to ensure improved model performance.

Keywords: Information overload, Information management, Natural language processing, Topic model, Latent dirichlet algorithm