УДК 004.9

A SELF-SUPERVISED LEARNING-BASED AUTOMATED BIBLIOGRAPHY ANALYSIS METHOD

Huiyao Dong (ITMO University)

Introduction. The volume and variety of research in COVID-19 have been steadily growing since the early years of the global pandemic. Researchers demonstrated that even in late 2020, the existing literature would already have covered a wide range of topics in medical and clinical research, society, mental health, supply-chain management, etc. [1]. To gain a thorough understanding of the recent research trend and identify the most valuable highlights from the massive literature, it is essential to learn and summarise the bibliometric features of existing research and their focus. With the help of deep learning-based text mining techniques, a comprehensive, interdisciplinary insight can be generated from the numerous papers on this crucial and long-lasting crisis. This paper proposes a self-supervised learning method for automatic bibliography analysis utilising BERT, an NLP representation technique developed by Google. According to the experiment, this method is proven to be capable of providing the COVID-19 papers' analysis in a robust manner, and the methodology can also be extended to the bibliography analysis of other topics or joint areas.

Body. The paper focuses on three main issues: (1) Implementing the BERT-based text mining methodology, including text vectorization and clustering, text embedding, and providing a comparative analysis between the BERT-based method and traditional text mining methods like Bag-of-words and Term Frequency-reverse Document Frequency (TF-RDF). (2) Developing an LSTM-based self-learning model, namely an auto encoder (AE), for sequential text data mining. (3) Displaying and presenting the result of the COVID-19-related biography analysis. BERT is an evolutional text mining model utilising attention mechanism to deduce the words' contextual relationships in input texts. Unlike previously proposed methods with a dedication to directional models, BERT allows bidirectional training with sequential data. It had two functional modules: an encoder for text processing and a decoder for target prediction; this paper only applies the encoder module as a part of the language processing mechanism. To generate the representation of multiple text-based features, this paper utilises the self-supervised learning approach, which can learn the characteristic behaviours and generate classification labels if needed without readily acquiring labels. The datasets used follow the features presented in an open dataset [3], but with a comprehensive collection of recent research papers until the end of 2022.

Findings. As demonstrated by the experiment result, BERT-based algorithm finishes the text mining tasks more efficiently, and the idea of utilising self-supervised learning is proven to be valid with sentence classification and clustering task. Besides, the analysis result of COVID-19 papers is displayed, in which topics include but not limited the research interests, methodologies applied and cross-subject researches. These results can be beneficial to both academic researchers and medical professionals.

Sources used:

1. Haghani M., Bliemer M., Goerlandt F., Li J. The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review // Safety Science. – 2020. – vol. 129. pp. 104806.
2. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint. – 2018.
3. Wang, Lucy Lu et al. CORD-19: The Covid-19 Open Research Dataset // ArXiv. – 2020.

Dong H. (author) Signature