2024 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 7-8 March 2024

Topic modelling of short texts in the health domain using LDA and bard

Mbooi, Mahlatse S
Council for Scientific and Industrial Research (CSIR)
Meiring Naude Drive, Pretoria, 0184
Email: MRatsoma@csir.co.za

This paper proposes a model for the topic modelling of tweets in the health and mental health domain using the Latent Dirichlet Allocation (LDA) method. The data were obtained from the sentiment140 project. The data were prepared for topic modelling by performing Natural Language Processing (NLP) tasks such as stemming and data cleaning. LDA method was trained on the data to create a cluster of topics. We explored 1 to 6 clusters and, after thorough analysis, three topics were chosen to create the LDA model. Each topic was labelled with a label name that is generated using Bard and coding analysis. This method can be used to label unlabelled data without using sophisticated supervised machine learning methods. Labelled data can be used to improve data management, information retrieval, supervised machine learning, and other techniques.