Abstract No: 267 ***ICT, Mathematics and Statistics***

ANALYSIS OF SOCIAL MEDIA CONTENT ON COVID CONTROL AND PREVENTION IN SRI LANKA

J.M. Hettiarachchi*

Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
**madarajayani@gmail.com*

Due to travel restrictions and stay-at-home orders brought about by the COVID-19 epidemic, social media platforms like Twitter have become an outlet for people to voice their worries, opinions, and feelings. Twitter is used by people, medical organizations, and governments to communicate about COVID-19. The goal of this study was to determine the effectiveness of COVID control and prevention in Sri Lanka by identifying the most discussed topics about COVID control and prevention among social media users and analyzing the sentiments of discussions on Twitter. The *Tweepy python library* and *Postman* with the Twitter *API* were used to collect English tweets related to COVID-19 control and prevention between January 1, 2020, and February 28, 2022. The analysis included topic modeling to identify topics and sentiment analysis to identify the emotions of Twitter users. The proposed methodology for topic modeling was based on the *coherence score*, which was used as the quality measurement for topic modeling. An *unsupervised machine learning algorithm (the Latent Dirichlet allocation algorithm)* was used to identify topics. For starting topic modeling, a combination of *unigram* and *bigram* was created to better understand key terms for identifying topics. For selecting better text feature vectors from the preprocessing dataset, both *BOW (Bag of Words)* and *TF-IDF (Term Frequency-Inverse Document Frequency)* were used for extracting features. The base model was trained with *TF-IDF* and with *BOW* and compared for the *coherence score*. The *hyperparameters* (number of topics, *alpha*) of a chosen *LDA* model were tuned to find the best combination of *hyperparameters* that results in the optimal number of topics. This study trained the final *LDA* model with the best *hyperparameter* combination and got the identified topics. *VADER (Valence Aware Dictionary and Sentiment Reasoner)* was used to identify the sentiments of each tweet. Each tweet was classified as positive, negative, or neutral based on the *compound score*. The results showed that the *coherence score* of the *LDA* model was increased through the proposed methodology. The increments in *coherence score* happened in three steps: the feature vector using *TF-IDF*, the optimal number of topics (k) = 21, and *alpha = "auto"*. Further, 21 topics were identified by combining *unigram* and *bigram* key terms, which resulted from topic modeling. Then, the results of sentiment analysis indicate 41% of tweets are negative, 33% positive, and 26% neutral. The main finding of this study is that Sri Lankans have negative feelings towards COVID control and prevention in Sri Lanka, as results of sentiment analysis indicate a higher percentage of total tweets are negative, and results of topic modeling indicate there are many negative topics being discussed by Twitter users. The *coherence score* is one quality measurement that can be used to evaluate topic modeling, and the combination of *unigram* and *bigram* provides better interpretability. Social media platforms such as Twitter are robust communication platforms that can be used to express public thoughts. The findings of the studies assist governments, the health sector, and the general public in raising awareness about the effectiveness of COVID control and in improving their strategies and plans for future health pandemics.

Keywords: Coherence score, Covid control and prevention, Sentiment analysis, Topic modeling, Tweets