Paper No: SC-15 **Smart Computing**

A Sentiment Analysis of COVID-19 Tweets Data Using Different Word Embedding Techniques

U.M.M.P.K. Nawarathne[1*], H.M.N.S. Kumari[2]

[1] *Computing Centre, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka, mnawarathne20@gmail.com*

[2] *Computing Centre, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka, nadeeshaku@sci.pdn.ac.lk*

The COVID-19 virus that invaded the world in 2019 caused many casualties while creating enormous mental turmoil among humans. During this pandemic period, humans were confined to prevent the virus from spreading. Due to the isolation, people used social media platforms like Twitter to express their ideas. Therefore, this study analyzed tweets related to COVID-19. Initially, text data processing techniques were employed, and sentiment labels were assigned. Then the data were trained using different machine learning (ML) models such as Multinomial Naïve Bayes (MNB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), k-Nearest Neighbours (KNN), Logistic Regression (LR), Extreme Gradient Boosting (XGB), and CatBoost (CB). During the training phase, word embedding techniques such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, Global Vectors for Word Representation (Glove), Bidirectional Encoder Representations from Transformers (BERT), and Robustly Optimized BERT-Pretraining Approach (RoBERTa) were used, and evaluation metrics such as accuracy, macro average precision, macro average recall, and macro average f1-score were calculated to evaluate these models. According to the results, the CB model, which used the RoBERTa technique, achieved an accuracy of 97%. Therefore, it can be concluded that CB with RoBERTa provides better results when classifying tweet data.

Keywords: *classification, machine learning, sentiment analysis, word embeddings*