Abstract No: SI-12

A deep learning approach to outbreak related tweet detection

B. A. S. S. B. Jayawardhana* and R. A. C. P. Rajapakse

Department of Industrial Management, Faculty of Science, University of Kelaniya, Sri Lanka
santhoopa@gmail.com*

Social Media has become a good indicator that reflects the real-time behaviour of society. Due to the popularity of social media platforms around the world, people use to express their observations and concerns on social media. People tend to report and discuss real-world events, personal health complications, and disaster situations through these platforms. These social media data streams can be used as a means to track and detect different types of events that affect large groups of people, such as epidemics, public disorderliness and disasters. Initial outbreak reports may first appear in these platforms even before it appears in the formal sources. A mechanism to identify these outbreak-related social media posts are needed to predict the outbreak in advance. Early detection of outbreaks in advance using these social media platforms will help relevant authorities to take appropriate actions. Even though there are existing models for outbreak prediction they have limited intelligence as they have focused only on one type of an outbreak. The main objective of this research is to propose a generalized model architecture that can detect tweets related to different types of outbreaks. In this paper, we propose a deep learning model that can detect tweets that are related to different outbreaks like epidemics, public disorders, and disasters. The semantic of the tweet is very important when determining whether it might be related to an outbreak. GloVe (Global Vectors for Word Representation) word embedding are used as the feature extraction technique in this study as it can capture the semantic meanings of the tweets. Long Short-Term Memory (LSTM) which is a specialized Recurrent Neural Network (RNN) architecture that can capture long-range dependencies in sequential data like text, is used as the classification algorithm. In the process, first, outbreak-related tweets were manually collected and labelled to ensure that only true outbreak-related tweets are fed into the supervised learning model. Then the annotated Twitter dataset of 4393 tweets was curated using relevant Natural Language Processing (NLP) techniques. Pre-trained GloVe word embedding of 100 dimensions that were trained on a large corpus of tweets were then used to represent the words of the tweets. As the next step, a Deep Learning Model was trained by using LSTM technique on the curated Twitter dataset. Finally, the performance of the model was evaluated using a different dataset of 341 tweets. During this phase, the model was evaluated using performance metrics, accuracy, precision, recall, and F1-score. The proposed deep learning model performed accurately in the testing dataset with an acceptable accuracy of 89%. The results were then compared with an existing machine learning model architecture for outbreak prediction. These results indicate the effectiveness of the LSTM algorithm when detecting outbreak-related tweets and the GloVe word embedding technique when capturing the semantics of tweets. With the results of this study, we can conclude that the proposed deep learning model architecture is an accurate approach for outbreak-related tweet detection.

Keywords: LSTM, NLP, Outbreak prediction, Twitter, Word embedding