# Leveraging Twitter data to Analyse the Virality of Covid-19 tweets: A Text Mining Approach

Geethu Joy*, Krishnadas Nanath

With the rapid growth of online social networks and media, understanding its influence on our thoughts and opinions are of prime importance. For an exceedingly long time, Twitter has been known to be used by government entities and organizations to understand trends and disseminate information. Covid-19 being the first pandemic in the current digital age that emerged with the outbreak and rapid spread of the novel coronavirus, in China, millions of users were seen discussing the pandemic on social media platforms. On 11th March, when Covid-19 was declared a pandemic by the WHO, the number of social media Covid mentions crossed 19 million, according to Sprinklr, the analytics platform. In early April, a report by Twitter said that COVID-19 tweets were being retweeted every 45 milliseconds. Being able to model and identify the factors that determine the virality of a tweet is necessary to prevent the spread of misinformation that can have dangerous consequences. This research takes up twitter data on Covid-19 to examine the virality of a tweet and attempts to develop a model that could help detect the retweet count of COVID-19 tweets.

We extracted tweets matching hashtags related to COVID-19 from March to May 2020, using Twitter's application programming interface. These tweets were subjected to big data techniques and filtered for the English Language, avoiding retweets as an entity, removal of missing value records. This resulted in a diverse collection of 47,330 tweets. We created new features such a hashtag, user mentions, popular hashtags, and length of tweets (using Python) from the extracted tweet data. The tweets were subjected to the following Natural Language Processing techniques: Topic modeling (LDA using Azure Machine Learning), Named Entity-Relationship (Azure ML), emotion, and sentiment analysis (using R Studio packages). This helped generate and test independent variables that could predict the retweet count and, thus, the virality of COVID-19 tweets. Log transformed dependent variable (retweet count), and the generated independent variables were subjected to linear regression. Our results indicated that tweets having named entities (person, organization, and location), frustrating emotions (anger and disgust), greater length, and inclusion of poplar hashtags had higher chances of being shared (retweeted). On the other hand, tweets having more number of hashtags and user mentions, anticipation/sadness emotion, would reduce the impact on retweets. These results could provide a perspective for brands to understand social media communication that could involve content related to the COVID-19 pandemic and position their posts accordingly.