Sentiment Analysis on Tweets from Genuine Twitter Accounts During the Covid-19 Pandemic

Sarthak Sengupta
Soumyadeep Basu
Anurika Vaish
Indian Institute of Information Technology- Allahabad, India

The COVID-19 pandemic has witnessed a massive surge in usage of social media and other online mediums for communication. It has resulted in a sudden increase in problems like fake news, cyberbullying, and targeted phishing attacks. A strategy to detect such issues and resolve them can be a boon for the users. This research study attempts to identify authentic Twitter accounts from bots during such trying times of crisis. Henceforth, the study devised the algorithm to perform sentiment analysis on the tweets from real Twitter accounts. The methodology and analysis of the study are based on data collected from the renowned microblogging social network platform named Twitter with the help programming language Python. Firstly, tweets are collected from bot accounts along with the non-bot accounts, i.e., genuine users based profiles. Then the algorithmic model was trained to differentiate between a bot and genuine accounts. This helped in the validation of genuine accounts. Various parameters were employed to define the differentiation rules like the time differences between various tweets posted from an account, follower versus following, usage of hashtags, no of retweets against actual tweets, etc. Nearly similar time gaps between two or more tweets increase the chances that the account under consideration is a bot. High and similar usage of hashtags across tweets and a high ratio of following versus followers are indicators of anomalous non-human behavior. Finally, tweets are scraped based on the keywords -COVID and India and Educational institutions. AND logic is used to take the intersection set of the keywords for definite time frames before and during the COVID-19 pandemic. The keywords-based dataset generated was then cross verified with the Twitter profile accounts data for testing if it is genuine or not. The fake or bot accounts along with their tweets were removed from the main dataset so that it contained tweets from genuine Twitter accounts only. After this step, sentiment analysis was done with the help of algorithms based on natural language processing. The algorithm had a data preprocessing stage which consisted of loading the data, cleaning the dataset, removing username, URL, numbers, emoji's special characters, etc. In the text processing stage, tokenization, stemming and lemmatization was done. Bag of words was used for feature extraction from the final tweet dataset. Various classification algorithms were tried and tested. Finally, the one based on support vector machine, and Xgboost was chosen. The research study finally performed the sentiment analysis on tweets from genuine Twitter accounts successfully after the model testing phase. The finally collected tweets were classified accordingly based on positive, negative, or neutral sentiments over brief periods. A general negative sentiment was observed that gradually decreased over time.

Keywords**:** *Genuine Profiles, Natural Language Processing, Python, Sentiment Analysis, Twitter*