A Sentiment Analysis of COVID-19 Tweets and its effects on Retweet-ability and Likeability.

**Andy Lee**  
500163559  
CIND820-D10

# Introduction

Social media has expanded tremendously in the past decade. It is common for government officials to use Twitter to relay public messages and announcements (Ontario Government, 2020), and also for individual users to rely on Twitter as their primary source of news (Glenski, et al., 2018). In the past, social media platforms has chosen a laissez-faire approach when dealing with user-generated messages (Caplan, 2017), however with a traffic of 500 millions tweets posted on Twitter per day (Pereira-Kohatsu, et al., 2019) where majority are posted by individuals users (Oren, et al., 2020), Twitter has slowly become a breeding ground for rumours and misinformation (Shao, et al., 2018); an ongoing issue social media platforms struggle to find a balance to address (Culliford, 2020).

As we step into 2020, the COVID-19 pandemic fills our everyday news while causing disturbance to all levels of society across the global (Hinshaw, 2020). Due to the novelty of the virus, governments and health officials often struggle to form straightforward and coherent guidelines and policies due to discovery of new information (Farzan, et al., 2020). This can contribute to public mistrust on the health authorities; indirectly fanning the spread of rumours and misinformation (Kouzy, et al., 2020). In turn, it hinders the effectiveness of public health policies when people feed on misinformation (Bode & Vraga, 2017) as studies have shown misinformation are spread and consumed by like-minded people; ricocheting like an echo chamber (Vicarioa, et al., 2016).

Monitoring the vast amounts of social media messages have become humanly impractical, therefore research into using machine learning and data analytic techniques to tackle the problem has become a popular research topic in recent years, and many papers have described partial or complete solutions with different degrees of success. In our capstone project, we will attempt to implement a small part of a big puzzle; using sentiment analysis techniques to label tweets on a scale of negative, positive, or neutral, and then build a classification model to predict whether sentimental levels can affect retweets or likes. This will allow us to better understand whether a tweet’s sentimentality can influence how it resonates with other people.

# Literature Review

**Twitter rumour detection in the health domain**

Add an introduction to create a smooth transition (1-2 sentences)

Sicilia et al. in “Twitter rumour detection in the health domain” (Sicilia, et al., 2018, pp. 34-35) described the construction of a complete detection system that can distinguish rumour or non-rumour tweets on topics related to the health sector. Rumours are defined as information from an unverified source, non-rumours are information that can be referenced to credible sources and official pages, and unknown are information that cannot be verified has either true or false (Sicilia, et al., 2018, p. 35). Data downloaded from Twitter, separated based on user or network level, feature selected based on performance, and finally trained into classification model (Sicilia, et al., 2018, pp. 35-36). The author used various machine learning techniques such as Support Vector Machine, Nearest Neighbour, and Random Forest and compared the results based on their averaged accuracies and compared their -values (Sicilia, et al., 2018, pp. 38-39). The study was able to achieve an over accuracy of 74%, precision of 73%, and recall of 74% (Sicilia, et al., 2018, p. 39).

Ravi in “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications” conducted a comprehensive review and experimented on different sentiment analysis techniques described in over 300 papers on six tasks, namely: subjectivity classification, sentiment classification, review usefulness measurement, lexicon creation, opinion spam detection, and aspect extraction” (Ravi, 2015, pp. 4-5). The study compared dozens of different sentiment analysis approaches and techniques with various degree of success on certain machine learning techniques when compared to vote count based measures (Ravi, 2015, pp. 55-57). Ravi concluded that there is still room for growth on intelligence-based techniques such as Random Forest (Ravi, 2015, p. 64).

Carlos et al. in “Detecting and Monitoring Hate Speech in Twitter” describes social media platforms dominate much of the internet, and with hundreds of millions of messages being conveyed through these platforms, hate messages begin to spread in wide varieties of subjects (Pereira-Kohatsu, et al., 2019, p. 2). The paper describes a system where it takes an input as text and emoji to help identifying and classifying hate speech in Twitter, and monitoring negative sentiments (Pereira-Kohatsu, et al., 2019, pp. 1-2). After experimenting with 19 different strategies of feature and classification models, the authors concluded with a model that combines LSTM and MLP-NN achieving AUC OF 0.828 (Pereira-Kohatsu, et al., 2019, p. 31).

Social media platforms currently are mostly unmoderated; allowing rumours or unverified information can easily spread and circulate on the platform (Zubiaga, et al., 2018, p. 32:1). Zubiaga et al. in “Detection and Resolution of Rumours in Social Media: A Survey” describes combining different various techniques in rumour detection, tracking, stance classification, and veracity classification together to form a rumour detection system that will provide users with early warning of messages containing uncertain information (Zubiaga, et al., 2018, pp. 32:1-32:2).

Sewalk et al. in “Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study” starts with the premise there is a shift to focusing more on patient experience and emotion, increasing transparency and patient engagement in health care (Sewalk, et al., 2018, p. 2). As Twitter has become a primary platform for people to voice their opinions, the authors conducted sentiment analysis on tweets based on patient tweets, such as wait time, as a method to evaluate their treatment experiences (Sewalk, et al., 2018, pp. 10-11).

# Dataset

For this project, we will use publicly available containing COVID-19 related tweets obtained from IEEE DataPort (<https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset>). The initial dataset contains only the tweets IDs filtered by a list of keywords ([link](https://rlamsal.com.np/keywords.tsv)) related to COVID-19. The tweet IDs is loaded onto a re-hydrator program to become tweets again. The resulting hydrated CSV have 34 attributes: coordinates, created\_at, hashtags, media, urls, favorite\_count, id, in\_reply\_to\_screen\_name, in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, lang, place, possibly\_sensitive, retweet\_count, retweet\_id, retweet\_screen\_name, source, text, tweet\_url, user\_created\_at, user\_screen\_name, user\_default\_profile\_image, user\_description, user\_favourites\_count, user\_followers\_count, user\_friends\_count, user\_listed\_count, user\_location, user\_name, user\_screen\_name, user\_statuses\_count, user\_time\_zone, user\_urls, user\_verified. For our project, the features selected are as follows:

Table 1 List of attributes used in this project.

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| created\_at | Time in UTC the tweet was created |
| favorite\_count | The number of “Likes” of the tweet |
| id | The id of the tweet |
| retweet\_count | The number of times this tweet has been retweeted |
| retweet\_id | The tweet id where the tweet was originally from |
| retweet\_screen\_name | The username of the person who originally wrote the tweet |
| source | The platform which the was written from |
| text | The context of the tweet |

# Approach

Figure 1 Project approach diagram

## Step 1: Download dataset from IEEE.org

Since the dataset is group by date, select a matching CSV date from IEEE.org. The study can be scaled and expand to multiple dates depending on computational power as some of the later processes are resource intensive.

## Step 2: Transform to text file

Import CSV file into Excel; ensure the tweet ID stays at integer and not converted into float by Excel. Remove the sentiment scores, leaving only one column of tweet IDs and save the file as a text document.

## Step 3: Load and run Hydrator

Download hydrator (<https://github.com/DocNow/hydrator>), load the tweet ID text document onto it, and begin rehydrating process which will take a few hours. Save the file as .CSV once hydrating process is done.

**Step 4: Preload Setup**

Load basic parameters such as full table widths to improve productivity while working on the project.

**Step 5: Import Dataset**

Store the previously hydrated file in some cloud storage services—we used Google Drive for this project.

To mount Google Drive onto the Colab project, first load the Google Colab library, then run and follow instructions to give Colab read privileges to the dataset file previously stored in Google Drive.

Once the drive is mounted, import CSV file onto Pandas dataframe.

**Step 6: Data Cleaning and Manipulation**

First observe and understand the dataset, then we can perform data cleaning such as handle null values, convert data types, and remove noise from strings.

Since retweets are duplicated tweets from an original tweet, we only need to focus on original tweets. Therefore, we filter the dataset to only include original tweets.

The algorithm works best on English text and tweets that are not in English will wastes resource and can affect our results. Using langdetect library, we parse each tweet text and discard any non-English data rows. This is a resource intensive task and require extended time to process.

**Step 7: Sentiment Analysis**

From NLTK, import SentimentIntensityAnalyzer and apply polarity score function onto each text inside the Original Tweet dataframe. The function returns positive, neutral, negative, and neutral in dictionary format, which is then expanded into their own columns inside the dataframe for better analysis.

Using the compound score, we can classify the tweet based on the range of scores.

**Step 8: Basic Analysis**

With all the features available, we can conduct some basic analysis such as correlation, mean, standard deviation, and histogram of the compound score.

**Step 9: Build Classification**

From Original Tweets, features: retweet\_count, user\_followers\_count, and compound will be our independent variables, and favorite\_count will be our dependent variable. The dataframe is broken into 80% training and 20% testing, and data will be fed into a linear regression model.

**Step 10: Getting Results**

Calculate metrics such as MSE, RMSE, and R-squared based on the predictions from classification models.

**Further Attempts**

We take the Original Tweets and break it into 80% training and 20% testing sets. Attributes retweet\_count, user\_followers\_count, compound will be our independent variables, and favorite\_count will be our dependent variable.

# References

Bode, L., & Vraga, E. K. (2017). See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication, 33*(9), 1131-1140. doi:10.1080/10410236.2017.1331312

Caplan, L. (2017, 10 11). *Should Facebook and Twitter Be Regulated Under the First Amendment?* Retrieved from WIRED: https://www.wired.com/story/should-facebook-and-twitter-be-regulated-under-the-first-amendment/

Carvalho, L. d., Silva, M. V., Costa, T. d., Oliveira, T. E., & Oliveira, G. A. (2020, 9). Public Health and the COVID-19 pandemic: challenges for global health. *Research, Society and Development, 9*(7). doi:10.33448/rsd-v9i7.4188

Culliford, E. (2020, 10 6). *Twitter is testing how its misinformation labels can be more obvious, direct*. Retrieved from Reuters: https://ca.reuters.com/article/us-usa-election-twitter-labels-focus-idCAKBN26R1ML

Farzan, A. N., Noack, R., Beachum, L., Taylor, A., Iati, M., Bellware, K., . . . Kornfield, M. (2020, 9 21). *CDC reverses statement on airborne transmission of coronavirus, says draft accidentally published*. Retrieved from Washington Post: https://www.washingtonpost.com/nation/2020/09/21/coronavirus-covid-live-updates-us/

Glenski, M., Weninger, T., & Volkova, S. (2018, 11 21). Propagation From Deceptive News Sources Who Shares, How Much, How Evenly, and How Quickly? *IEEE Xplore*, 1071-1082. doi:10.1109/TCSS.2018.2881071

Hinshaw, D. (2020, 10 12). *As Covid Cases Surge, More Public-Health Experts Say Lockdowns Aren’t the Answer*. Retrieved from Wall Street Journal: https://www.wsj.com/articles/public-health-experts-rethink-lockdowns-as-covid-cases-surge-11602514769

Kouzy, R., Jaoude, J. A., Kraitem, A., Alam, M. B., Karam, B., Adib, E., . . . Baddour, K. (2020, 3 13). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*. doi:10.7759/cureus.7255

Ontario Government. (2020, 10 1). *Official Government of Ontario Twitter*. Retrieved from Twitter: https://twitter.com/ongov

Oren, E., Martinez, L., Hensley, R. E., Jain, P., Ahmed, T., Purnajo, I., . . . Tsou, M.-H. (2020, 7 29). Twitter Communication During an Outbreak of Hepatitis A in San Diego, 2016–2018. *American Journal of Public Health*. doi:10.2105/AJPH.2020.305900

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019, 11 19). Detecting and Monitoring Hate Speech in Twitter. *sensors, 19*(21), 4654. doi:10.3390/s19214654

Ravi, K. (2015, 6). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based System, 89*, 14-46. doi:10.1016/j.knosys.2015.06.015

Sewalk, K. C., Tuli, G., Hswen, Y., Brownstein, J. S., & Hawkins, J. B. (2018, 10). Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study. *Journal of Medical Internet Research, 20*(10). doi:10.2196/10043

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018, 4). Anatomy of an online misinformation network. *PLoS One, 13*(4). doi:10.1371/journal.pone.0196087

Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018, 11 15). Twitter rumour detection in the health domain. *Expert Systems With Applications, 110*, 33-40. doi:10.1016/j.eswa.2018.05.019

Vicarioa, M. D., Bessib, A., Zolloa, F., Petronic, F., Scalaa, A., Caldarellia, G., . . . Quattrociocchia, W. (2016, 1 19). The spreading of misinformation online. *Proceedings of the National Academy of Sciences - PNAS, 113*(3), 554-9. doi:10.1073/pnas.1517441113

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, a. R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys, 51*(2). doi:10.1145/3161603