Andrea Chung
Fall 2022

**Sentiment Analysis on Outbreak of Covid-19 in USA and its Changes over time (1345 words)**

## Introduction

Social media platforms such as Twitter has been suitable grounds for mining public's opinion on issues from today's society. The data scraped from Twitter can be used to perform sentiment analysis to follow people's attitude toward certain matters. Coronavirus (Covid-19) has become one of the significant disease outbreaks, which is continued until today. Because the virus did not have appropriate treatments, the spread and increase in its severity were inevitable. Also, quarantine and keeping social distance were necessary, which caused Twitter to emerge as one of the platforms where people can express their views and eagerness (Singh et al., 2021).

## Background/Related Study

During Covid-19, Twitter was used in various ways. Since the nature of online platforms allows people to share information faster than books or journals, the amount of data we can achieve is incomparable (Garcia & Berton, 2021). Especially, hashtag tools in Twitter facilitate greatly in obtaining related tweets and observing user's sentiment trends accordingly (Garcia & Berton, 2021). As one of examples, studied research was on sentiment analysis of tweets on Covid-19 using bidirectional encoder representation from transformers (BERT) as classifier technique (Singh et al., 2021). Different measurement methods were used such as "average likes over period, average retweets over the period, intensity analysis, polarity, subjectivity, and Wordcloud" (Singh et al., 2021). Additionally, Singh et al. compared the analysis between data from people from the world and people from specific country, India. The result obtained from the model had validation accuracy of 94%. The major findings of the paper were that people from India comparatively had more positivity, which they made correlation between the sentiments and "success or failure of the measures adopted by government of a country in various circumstances (Singh et al., 2021)."

Another research utilizes long short-term memory (LSTM) to conduct sentiment analysis on social media data to observe Covid-19 opinions and forecast on number of cases. They make comparison between Textblob model and LSTM and retrieve that performance of LSTM model was greater with accuracy and efficiency than Textblob model (Alorini et al., 2021). The major finding of this journal was that with the prediction made from LSTM model, they forecasted number of Covid-19 cases, which may be used to alert publics ahead (Alorini et al., 2021).

Through this paper, I intend to analyze public's sentiment and its changes over time on "quarantine" from the beginning of Covid-19 hit on USA and over the 6 months period. The quarantine was inevitable during the period so, I wanted to investigate how people's opinion changed over time and answer the following research question:

- How does public's sentiment change towards "quarantine" between the start of Covid-19 and after six months of the outbreak?

## Method

*Data*

To answer the research question above, I scraped tweets from 6 different timeframes, which are between March 2020 and August 2020. I chose March 2020 because it was when Covid-19 cases started to increase enormously in USA. 1000 tweets were collected from each month resulting in around 6000 tweets in total. The data was collected using "snscrape" package to avoid the rate limit and time-period limit. To implement TwitterSearchScraper function, following keywords and hashtags were used: "covid19", "quarantine", "#covid19", and "quarantine". All the duplicated tweets were deleted, and tweets only written

in English was collected. Because the Covid-19 was declared as pandemic and disseminated over the world, I did not specify the geo-location of the scraped tweets.

*Sentiment Analysis*

The sentiment analysis of the tweets was done by creating and performing deep learning model using LSTM. LSTM is one of the newly developed recurrent neural networks (RNN) model and is often used for different analysis in natural language processing (NLP) areas. When implementing sentiment analysis on RNN, there are drawbacks, and the model becomes unstable when there is large amount of information going through which results in enormous updates (Srivastava, 2020). However, LSTM helps to overcome these problems by using gates to manage the memorizing process. Long-term "memory" makes the model more appropriate and capable of understanding the overall context better than other neural networks affected by long-term dependency problem (Srivastava, 2020). LSTM works with interaction of two types of gates and are used to choose which information will pass through or not. The sigmoid function gives output of 0 or 1, which the value 0 makes data not to pass through when value 1 permits all the information to pass through the first gate (Srivastava, 2020).

Before applying LSTM model, preprocessing and tokenization was done on both training data texts and scraped tweets data of mine. Unnecessary words such as urls, htmls, @mentions, hashtag signs, "\n", and English stop-words were removed from the text to increase the efficiency and accuracy level of analysis. Also, all the texts were lower cased. For this study, pre-existing training data set, Sentiment140, was used to train my model and evaluated the model by the accuracy level. Sentiment140 is one of the favored data to be used as training data, which has 160,000 tweets collected using Twitter API by researchers at Stanford University (Sentiment140). Using train_test_split module, the 80% of the data was used for training and 20% of the data was used for testing.

For word embedding, pre-trained word embedding document, Global Vectors for Word Representation (GloVe), was used. Particularly, 200-dimensional word embedding of tweets were used provided by Pennington et. al (Pennington et al., 2014). The following image shows the final model structure (Image 1).

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 52, 200)           51250400

dropout (Dropout)            (None, 52, 200)           0

lstm (LSTM)                  (None, 200)               320800

dense (Dense)                (None, 64)                12864

dense_1 (Dense)              (None, 1)                 65

=================================================================
Total params: 51,584,129
Trainable params: 333,729
Non-trainable params: 51,250,400
_________________________________________________________________
```

Image 1. Model Structure

The dropout layer was added to prevent the overfitting of 20% of neurons. The model had training accuracy of 85.69% and testing accuracy of 84.38%. The model was used with 6 datasets I have acquired from Twitter. The tweets were labelled with only two labels, positive and negative by the score calculated from the model for each text. The following table shows example of resulting output (Table 1).

| | Text | Score | Label |
|---|---|---|---|
| 0 | thanks just the extra scoop of depression i ne... | 0.790711 | Positive |
| 1 | quarantine realization it is impossible to wo... | 0.060923 | Negative |
| 2 | it s not fair snapchat socialdistancing quaran... | 0.054282 | Negative |
| 3 | since this quarantine started the risk of star... | 0.074786 | Negative |
| 4 | thank u our police and doctors stayhomesaveliv... | 0.600242 | Positive |

Table 1. Example of Labelled Tweets with Score

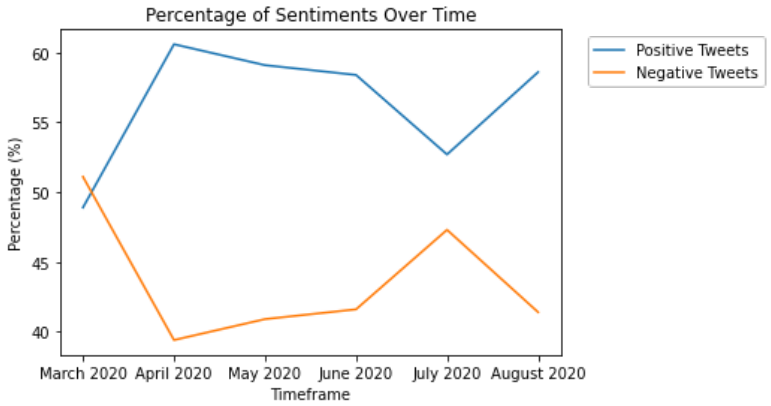
**Result**

The following table shows the distribution of positive and negative tweets for each timeframe (Table 2). The percentile distribution was used because the number of tweets for each month was different after removing duplicates.

| | Timeframe | Percentage of Positive Tweets | Percentage of Negative Tweets |
|---|---|---|---|
| 0 | March 2020 | 48.9 | 51.1 |
| 1 | April 2020 | 60.6 | 39.4 |
| 2 | May 2020 | 59.1 | 40.9 |
| 3 | June 2020 | 58.4 | 41.6 |
| 4 | July 2020 | 52.7 | 47.3 |
| 5 | August 2020 | 58.6 | 41.4 |

Table 2. Distribution of Positive and Negative Tweets over Time

Looking at the table, except in March 2020, the percentage of positive tweets was higher than percentage of negative tweets. From this we can speculate that because March was the starting month of the new pandemic hit on USA and on most of other countries around the world, people tend to have more negative emotions expressed with new adjustments through quarantine and lockdowns from many countries. However, from April 2020, percentage of positive tweets tend to be higher, but the number drops slightly on July 2020. Moreover, we can visually observe the distribution through following plot:




Looking at the graph above, we can notice that there is large dip on the trend line of percentage of positive tweets in July 2020. The fluctuation somewhat shows the instability of society due to unpredicted enlargement of the disease around the world and inevitable physical quarantines.

**Conclusion and Limitation**

In conclusion, people's sentiment changes over time greatly on societal acts such as quarantine and social distancing due to outbreak of Covid-19. With the deep learning model using LSTM, I was able to perform sentiment analysis with relatively desirable accuracy level. Through the analysis above, we can conclude that public's opinion fluctuates but positivity remains higher than negativity toward quarantine during the outbreak of the disease.

For further study, the analysis can be done over longer period with a greater number of tweets. Since the pandemic has been continued until today, higher number of tweets can be collected for each month over 2 years span. Moreover, another machine learning method can be used such as convolutional neural network to create another sentiment analysis model and compare on the performance to see which method is more appropriate. Also, sentiment analysis can be done by different countries or regions using geo-location as a part of the extension to the current research.

**Reference**

Alorini, G., Rawat, D. B., & Alorini, D. (2021). LSTM-RNN based sentiment analysis to monitor covid-19 opinions using social media data. *ICC 2021 - IEEE International Conference on Communications*. https://doi.org/10.1109/icc42927.2021.9500897

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to covid-19 from Brazil and the USA. *Applied Soft Computing*, *101*, 107057. https://doi.org/10.1016/j.asoc.2020.107057

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/d14-1162

Singh, M., Jakhar, A. K., & Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the bert model. *Social Network Analysis and Mining*, *11*(1). https://doi.org/10.1007/s13278-021-00737-z

Srivastava, P. (2020, May 18). *Long short term memory: Architecture of LSTM*. Analytics Vidhya. Retrieved December 8, 2022, from https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/

*A Twitter sentiment analysis tool*. Sentiment140. (n.d.). Retrieved December 8, 2022, from http://help.sentiment140.com/home