Modern Applied Science; Vol. 14, No. 11; 2020
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Online Messages Sentiments Analysis Based on Long Short-Term Memory

Yunke Zhao[1]

[1] Shanghai Foreign Language School, China

Correspondence: Yunke Zhao, Shanghai Foreign Language School, China.

Received: September 11, 2020 Accepted: October 19, 2020 Online Published: October 28, 2020

doi:10.5539/mas.v14n11p36 URL: https://doi.org/10.5539/mas.v14n11p36

Abstract

In December of 2019, an extremely infectious and deadly pandemic ambushed China. In Wuhan, the novel coronavirus COVID-19 suddenly broke out and spread rapidly to other countries. COVID-19 became a worldwide disaster, affecting not only physical, but also emotional health on a global scale. We wanted to record this change based on the sentiment analysis model and to examine the relationship between world events and the positivity of posts on social media.

To analyze this relationship, we utilized a set of movie reviews as a training sample to construct a sentiment analysis model based on the Long Short-Term Memory neural network theory, and calculate the texts' sentiment score. We then analyzed the overall trend of the data, and discussed the reason behind the tendency. The principal result was that, as the pandemic progressed, online sentiment generally became more positive. We believe that this is because people gradually become more accustomed to life in the COVID-19 era.

Keywords: coronavirus, long short-term memory, neural network, sentiment analysis

1. Introduction

COVID-19, which broke out in 2019 and 2020, has caused tremendous loss worldwide. This kind of virus is not only extremely infectious, but it is also quite deadly, with a long incubation period and a permanently high reproduction number, which is hard to decrease. As of August 24, 2020, there are already 23,311,719 cumulative confirmed cases worldwide, and that number will grow continually. Nevertheless, with the guidance of the WHO and the efforts of governments, experts, and individuals, the virus is gradually becoming controlled. Only 44,946 new cases are being confirmed in the United States on August 24, significantly less than the 69,641 confirmed cases on July 24. (Note 1) That number is around 50,000 on October 11.

During the pandemic, various events have triggered different responses on the Internet, leading to social media users expressing different sentiments during different periods. With little idea of the scale of the pandemic, initially, the public may have held a neutral position. As the pandemic progressed, there might be a more pessimistic response; and conversely, the public may become somewhat hopeful and optimistic again once they have gotten used to the new life and believe that the pandemic will soon be resolved. Hence, the article will research the sentimental change of the messages on social media based on LSTM and analyze the results.

This article will include the following approaches:

1) The data retrieved from movie reviews will be used to train a sentiment analysis model based on a long short-term memory neural network. This model should automatically classify passages as positive, negative, or neutral based on the experience in the sample training.

2) The sentiment analysis model, will be used to calculate sentiment scores and emotion tags for messages extracted from social media during the pandemic period.

3) Use Gephi and other means will be used to visualize the data and study the pattern of the result. The reasons behind the trend will be analyzed by relating sentiment scores to events in reality.

2. Assumptions

1) In consideration of the limitations of data collection, the article assumes that the messages in the news/message boards/blogs during December in 2019 to March in 2020 and tweets during April to July in 2020

are representative of the public's overall messages on social media and can validly express the public's sentiment.

2) Given that only English texts are gathered during the data collection, the problems this paper discussed are valid only in the United States, Britain, and other English-speaking countries or districts. Due to the lack of the data source location, the article assumes no significant difference among the data from different areas like urban cities and rural areas.

3) In consideration of the fact that the article's sentiment analysis model calculates only messages from December 2019 to March 2020, the article assumes that the TextBlob model used for the ready-calculated sentiment scores of tweets from April to July has roughly similar criteria for the score as the article's sentiment analysis model's.

4) During the calculation of the sentiment scores for the December-to-March texts, only randomly selected 1% of the data are used, and the article assumes that this selected part is representative of the whole population.

3. Data Collection

3.1 Movie Review Data Set

The movie review dataset serves as a sample and test data to train the article's sentiment analysis model based on the LSTM. Downloaded from GitHub open-source (https://github.com/yangbeans /Text_Classification_LSTM), the dataset has two files containing positive and negative film reviews, correspondingly.

C++ was used to select and randomly order, and we divide the data into 431 positive files and 428 negative files, each containing around 300 words and ending with a full sentence. To make sure that the machine will not randomly guess the text's sentiment tag based on the proportion of the sample, we equally assign 400 positive files and 400 negative files to the sample, and the rest of the data will be the test data.

3.2 News/Message Boards/Blogs Data Set

The news/message boards/blogs data set is the subject of the sentiment analysis. The texts extracted from this data set will be used for sentiment score calculation and will be evaluated as positive or negative. Downloaded from IEEE Dataport, (Note 2) the data set contains messages on social media from December 2019 to March 2020.

Using a python program, we extracted the texts from JSON files and divided them into text files, each containing one single message labeled with the date. Then we gathered all the December data and randomly selected 1% of the January-to-March data in each day as our data files, each containing the date label and the message texts. Most of the texts are around 200 words, and the majority of the texts have a word amount less than 300, fit for the article's sentiment analysis model.

Using the sentiment analysis model, we obtained the sentiment score for each message and built a news/message boards/blogs data set, each containing the date tag and the corresponding sentiment score.

3.3 Corona-Virus (COVID-19) Geo-Tagged Tweets Data Set

The coronavirus (COVID-19) geo-tagged tweets data set and the worked December-to-March data set serve as the subject of sentiment trend analysis. This data set is downloaded from IEEE Dataport (Note 3) and it contains the tweet IDs and the corresponding sentiment scores from April to July. The sentiment score inside the data set is calculated by TextBlob's simplified text processing model, used for sentiment analysis of emotion polarity, roughly the same criteria as our model's.

We gathered the data set's sentiment score, arranged the data, and labeled its distribution and proportion of different sentiments to create the data for the period from April to July.

4. Basic Theory about Neural Network

4.1 Definition of Neural Network

The artificial neural network is a kind of calculation model mimicking the biological neural network using mathematical and computer science arithmetic. The neural network structure includes multiple layers of artificial neurons, and the model calculates by transmitting the data and result through layers. Like the biological neural network, an artificial neural network can learn and adjust its parameters and structure to minimize the assigned loss according to the feedback each time. The model can be used to solve regression problems, do image recognition, analyze text, etc.

The neural network mainly has two strategies of study: supervised learning and unsupervised learning.

Supervised learning lets the machine observe the pre-labeled sample data by humans and find the pattern. The machine will adjust its parameter and structure to form a model to predict the output result when the input is different from the sample. The supervised learning strategy requires a large amount of data, allowing the machine's sample data to learn and the test data for the human to determine the model's efficiency. Supervised learning can solve regression problems as well as sentiment analysis of the text.

Unsupervised learning, on the other hand, lets the machine automatically divide the non-pre-labeled sample data into classification. The model can be employed for clustering analysis and Generative Adversarial Networks.

4.2 Definition of RNN and LSTM

RNN, Recurrent neural network, is a kind of artificial neural network typically used for automatic speech recognition and natural language processing. Instead of merely processing each data as the typical neural network does, the feedback strategy allows RNN to detect and learn the pattern in a sequence by reprocessing some data. This tactic can be beneficial when it comes to the field of natural language processing. Take machine translation, for example: word-to-word translation is a horrible strategy, and sometimes the meaning of the word is dependent on the context in that sentence.RNN can analyze the whole sequence and deal with that pattern. For this reason, RNN is extremely powerful when the context of the sequence is critical for analyzing the data.

LSTM, Long short-term memory, is a typical recurrent neural network architecture. Its feedback connections allow the model to analyze the sequence comprehensively. Long short-term memory model usually contains several components including a cell and three gates—input gate, output gate, and forget gate—by which the machine can decide which information should be remembered for further calculation and which should be discarded to reduce the amount of memory. The selective forgetting and memorizing function of the Long short-term memory avoids vanishing and exploding gradients, which could cause severe problems in a standard recurrent neural network structure because of RNN's memorize-all strategy. Hence LSTM can perform well when dealing with sentiment analysis in which information from the past data will be needed to analyze new data in the sequence.

5. Sentiment Analysis Model Based on LSTM

5.1 Model Introduction and Efficiency

The objective is to use a long short-term memory model to analyze the sentiment in a text. The model employs the 100-dimension Glove Vector by Stanford for the vocabulary. With the assigned sample data and the December-to-March messages' test data, the model reads the sample text and automatically adjusts the word amount of each input data to 300 by cutting off longer messages, and complementing the shorter one with zeros and process the text with pre-trained vocabulary. In consideration of the amount of sample and other factors, we chose a batch size of 40 and set the learning rates as 0.01 and the number of epochs as 20. The calculation result will be measured by a sentiment score, varied from -1 to 1. The negative index indicates a negative emotion, and a positive one indicates positive emotion. The absolute value of the score indicates polarity. The sentiment is more extreme when the absolute value of the score is close to 1, and the position tends to be neutral if the score is around zero.

The model runs for 20 epochs, and the sample accuracy increases from around 50% to over 96%. Simultaneously, the analysis of the test data, which includes 31 positive files and 28 negative files, manifests a high efficiency of the model with an accuracy close to 95%. It can be concluded that the model is pretty efficient and confidently valid.

The core part of the model’s code is as Figure 1

```
class BiRNN(nn.Module):
    def_init_(self,vocab,embed_size,num_hiddens,num_layers):
        super(BiRNN,self)._init_()
        self.embedding = nn.Embedding(len(vocab),embed_size)
        self.encoder = nn.LSTM(input_size = embed_size,
                               hidden_size = num_hiddens,
                               num_layers = num_layers,
                               bidirectional = True)
        self.decoder = nn.Linear(4*num_hiddens,2)
    def forward(self,inputs):
        embeddings = self.embedding(inputs.permute(1.0))
        outputs, _= self.encoder(embeddings)
        encoding = torch.cat((outputs[0],outputs[-1]),-1)
        outs = self.decoder(encoding)
        return outs
        ac_tra = 0
        ac_tes = 0
        embed_size,num_hiddens,num_layers = 200,300,2
        net = BiRNN(vocab,embed_size,num_hiddens,num_layers)
        glove_vocab = Vocab.GloVe(name = '6B',dim = 200,cache = os.path.join(DATA_ROOT,
        "glove"))
        lr,num_epochs = 0.01,20
        optimizer = torch.optim.Adam(filter(lambda p:p.requires_grad,net.parameters()),lr=lr)
        loss = nn.CrossEntropyLoss()
        d21.train(train_iter,test_iter,net,loss,optimizer,device,num_epochs)
```

Figure 1. Model Code

5.2 The Result of the Sentiment Analysis

We collate the messages' sentiment score results from December to March and integrate them with the pre-calculated sentiment score results of the tweet data from April to July into a single sentiment distribution file. Then we define a more specific sentiment tag. We determine that a data with sentiment score greater than or equal to 0.5 is considered "positive," and one with that between 0.1 and 0.5, exclusively, is considered to be "neutral towards positive." Similarly, the message with a score smaller than or equal to -0.5 is "negative," while text with a score between -0.5 and -0.1, exclusively, is "neutral towards negative." The data with a sentiment score between -0.1 and 0.1, inclusively, is determined to be "neutral."

By assigning the sentiment tag to each message and counting each tag's appearance number in a single day, we get the proportion of the five sentiment tags in each day. Considering the lack of data in December and frequent fluctuation of the data, we merge the data by weeks. The first seven days in a month will be referred to as the first week of that month. The messages in a month will be divided into four files, each of which contains the average proportion of different sentiment tags in the corresponding week while the data of the last remaining days beyond four weeks in the month is merged with the 4th week's data. Finally, we get the proportion of sentiment tags in each week unit, shown as Figure 2:

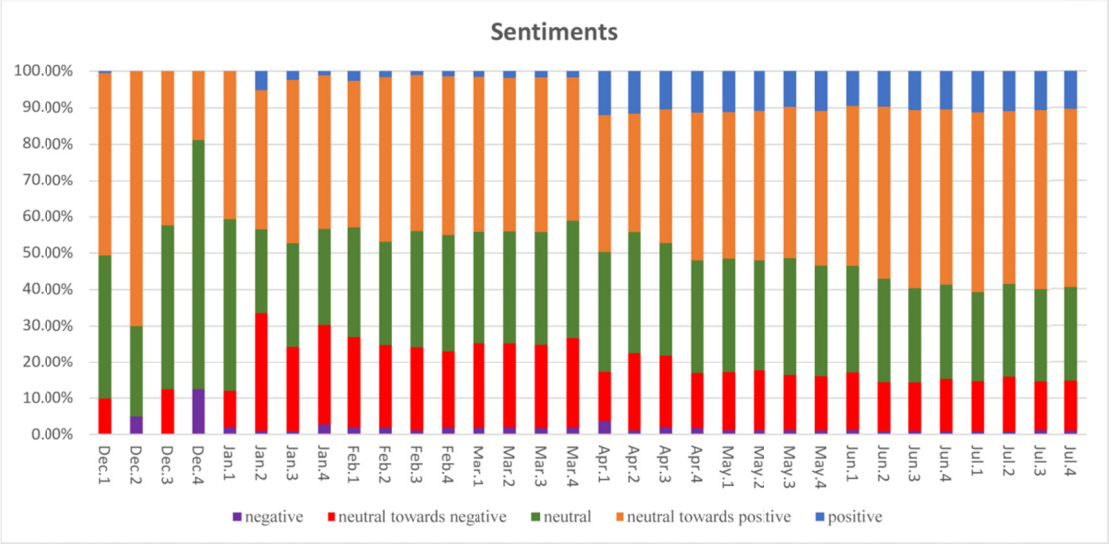

Figure 2. Histogram of Sentiment Proportion in Each Time Period

In Figure 2, the x-axis of the graph represents the periods of the data, where April 1 means the 1st week in April, and July 4 means the 4th week in July plus July 29th, 30th, and 31st. The y-axis stands for the proportion in the messages. The color indicates the sentiment type, as is described at the bottom of the graph.

6. Analysis of the Result

6.1 Discussion of the Overall Pattern

6.1.1 Evaluation of the Result and Variation

To analyze the overall trend of the sentiment proportion, we draw a line graph in which each line represents a sentiment tag, as is shown as Figure 3:

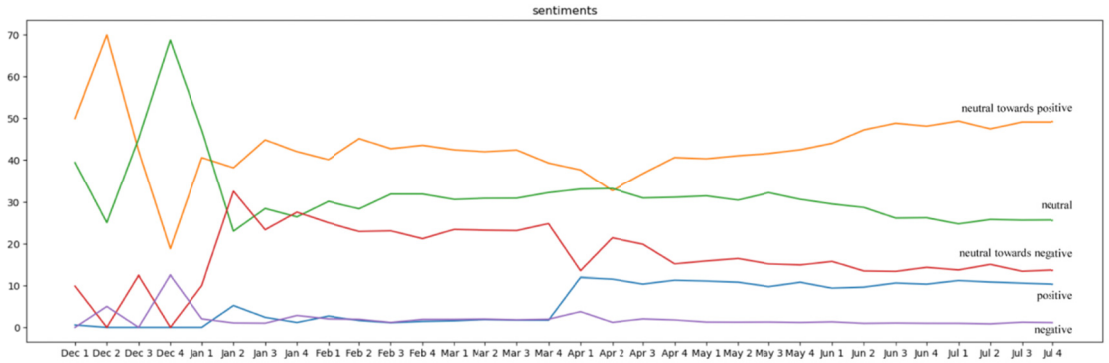

Figure 3. Line Graph of Sentiment Proportion in Each Time Period

In Figure 3, the x-axis of the graph represents the time period of the data, the same as the histogram. The graph's y-axis indicates the percent of the certain sentiment tag in the corresponding period with a unit of 1%. The color of the line indicates the type of sentiment, the same as that in the histogram. The blue line indicates "positive," orange is for "neutral towards positive," the green one is "neutral," the red line is for "neutral towards negative," and the purple represents "negative." The score before April 1, exclusively, is calculated by the article's sentiment analysis model, and the data after April, inclusively, is gathered from the open-source data calculated by the TextBlob Model. Though the data are from two different data sets, we can see the lines are relatively smooth, and hence we can conclude that the joint of the two data is quite acceptable.

From the line graph, we can see a massive fluctuation of proportion in the first five periods, and the line becomes smoother when the time is later. This is due to the small data size in December, while partial texts even refer to the SARS virus in 2003. Hence the fluctuation of the initial data is gigantic. However, the variation tends to decrease as the outbreak spreads worldwide as the number of messages regarding COVID-19 starts to skyrocket.

6.1.2 Discussion of the "Positive" Sentiment

By using A regression model based on the neural network, we examine the data of the "positive" sentiment,

excluding those of the first five periods of weeks which are fluctuating, and calculate the best-fit regression function, which can be expressed as $y = -67.9508x^2 + 69.2357x - 3.2391$, where x is the number of periods of weeks divided by 100, drawn as Figure 4:

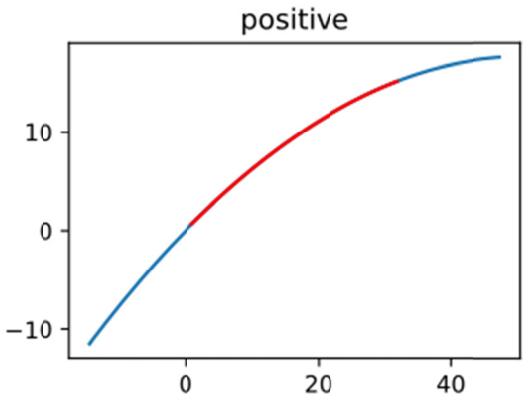

Figure 4. Regression Function of Positive Sentiment

In Figure 4, the x-axis is the periods of weeks from the first week of December, while December 1st to 7th will be counted as 1 and December 22tnd to 31st will be 4. The y-axis is the predicted percentage of the "positive" sentiment with a unit of 1%. The red part represents the prediction of the function in the definition domain of the December to July data, and the blue part is for all valid domain of definition.

The graph for the actual function and comparison between it and the predicted function is as Figure 5 and Figure 6:

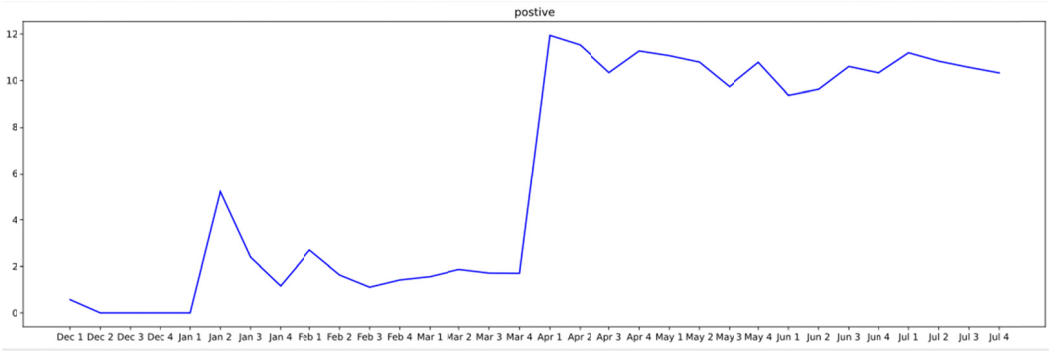

Figure 5. Actual Function of Positive Sentiment

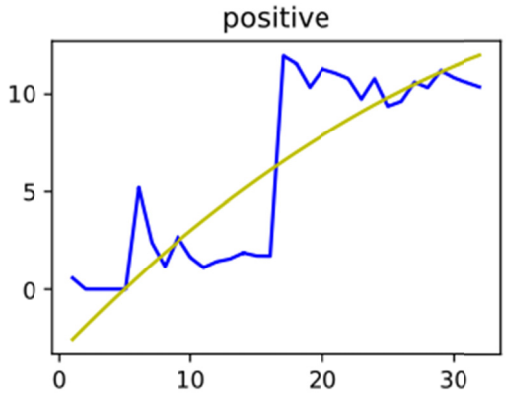

Figure 6. Actual vs Regression Function of Positive Sentiment

Figure 5 represents the actual function, and the second is the comparison between the actual and the predicted. In the first graph, the x-axis is the period, written by the week's number in a month. The y-axis represents the percentage of that sentiment in all messages. In Figure 6, the x-axis is the periods of weeks from the first week of December, the same as the graph mentioned earlier. The y-axis is the percentage of the "positive" sentiment,

while the blue line represents the actual percentage, and the yellow line indicates the predicted proportion.

It can be seen from the Figure 4 that the trend of "positive" percentage shows extreme linearity. That is to say: generally, more people will manifest positive emotion in their message as the pandemic progresses. This logic seems unreasonable on the surface, but in fact, it can be explained by the conjecture that, as the pandemic progresses, people are gradually becoming used to their new lives in the era of coronavirus and have stopped panicking as they were initially. Quarantine and inconvenience may not seem as horrible as they appeared at the beginning of the pandemic. Besides, the fact that governments and experts might continue to spread positive news and messages to pacify and mollify people may also trigger more positive responses. Of course, since the "positive" messages partake little in the whole data, these differences might be minor in consideration of the whole picture. Also, the data sets change seems to trigger a tremendous change in the graph between March 4th and April 1st, making the trends less valid.

6.1.3 Discussion of the "Positive Towards Neutral" Sentiment

By using regression model based on neural network, we examine the data of the "neutral towards positive" sentiment, excluding those of the first five periods of weeks which are fluctuating, and calculate the best-fit regression function, which can be expressed as $y = 249.7752x^2 - 65.1676x + 44.6142$, where x is the number of periods of weeks divided by 100, drawn as Figure 7:

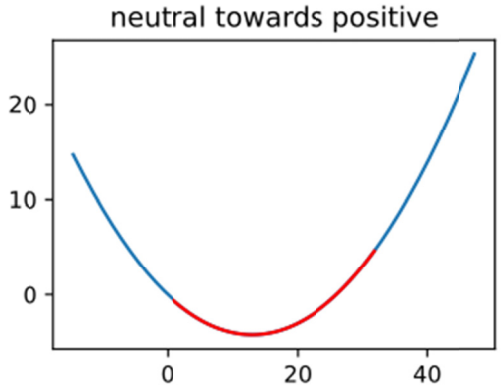

Figure7. Regression Function of Neutral towards Positive Sentiment

In Figure 7, the x-axis is the period number from the first week of December, the same as the graph mentioned earlier. The y-axis is the predicted percentage of the "neutral towards positive" sentiment with a unit of 1%. The red part represents the prediction of the function in the definition domain of the December to July data, and the blue part is for all valid domain of definition.

The graph for the actual function and comparison between it and the predicted function is as Figure 8 and Figure 9:

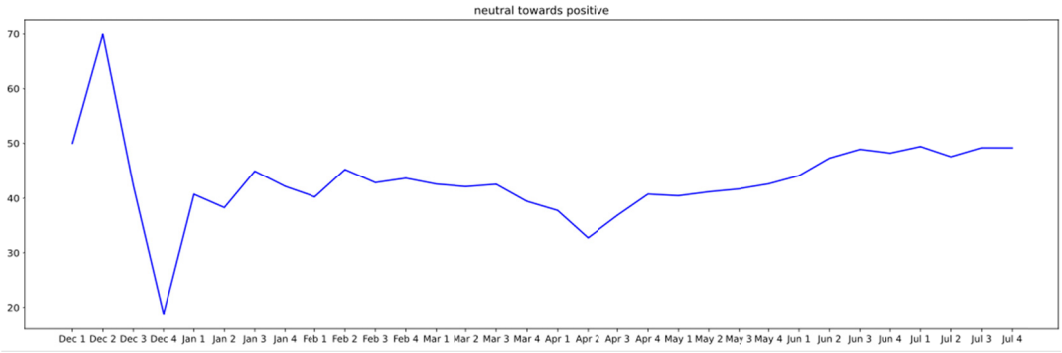

Figure 8. Actual Function of Neutral towards Positive Sentiment

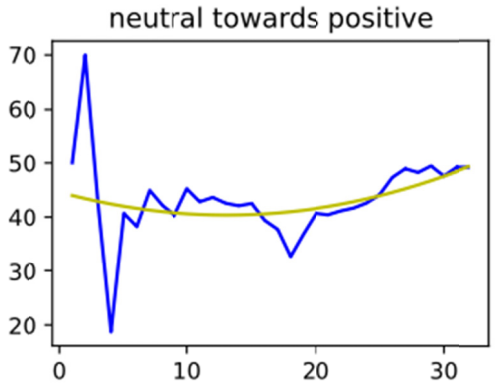

Figure 9. Actual vs Regression Function of Neutral towards Positive Sentiment

Figure 8 represents the actual function, and the second is the comparison between the actual and the predicted. In the first graph, the x-axis is the period, written by the week's number in a month. The y-axis represents the percentage of that sentiment in all messages. In Figure 9, the x-axis is the periods of weeks from the first week of December, the same as the graph mentioned earlier. The y-axis is the percentage of the "neutral towards positive" sentiment, while the blue line represents the actual percentage, and the yellow line indicates the predicted proportion.

It can be seen from Figure 8 and Figure 9 that, besides the first several data in December with small data size and a vast variation, generally the "neutral towards positive" sentiment has a roughly constant pattern at the beginning of the pandemic and a gradually growing trend in the latter part of the coronavirus period. It can be known that from December to March, people are less positive in their messages, and this can be interpreted as the result of the outbreak of the virus. Since the virus has proven extremely infectious and has spread to many countries, and since people indeed will not have had time to become used to a new life with many restrictions because of the virus (including quarantine and social distancing), the public will seem more pessimistic and hopeless, thus having fewer positive messages. Later, people express more positivity when they have gradually become used to their new life, and as the global situation improves (e.g. as China sees zero daily new confirmed cases since the beginning of April).

6.1.4 Discussion of the “Neutral” Sentiment

By using regression model based on neural network, we examine the data of the “neutral” sentiment, excluding those of the first five periods of weeks which are fluctuating, and calculate the best-fit regression function, which can be expressed as $y = -261.0937x^2 + 89.0208x + 23.8949$, where x is the number of periods of weeks divided by 100, drawn as Figure 10:

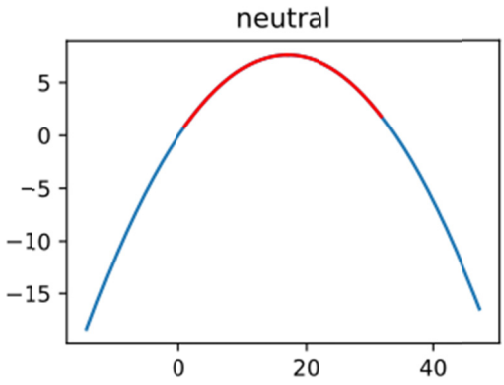

Figure 10. Regression Function of Neutral Sentiment

In Figure 10, the x-axis is the period number from the first week of December, the same as the graph mentioned earlier. The y-axis is the predicted percentage of the "neutral" sentiment with a unit of 1%. The red part represents the prediction of the function in the definition domain of the December to July data, and the blue part is for all valid domain of definition.

The graph for the actual function and comparison between it and the predicted function is as Figure 11 and Figure 12:

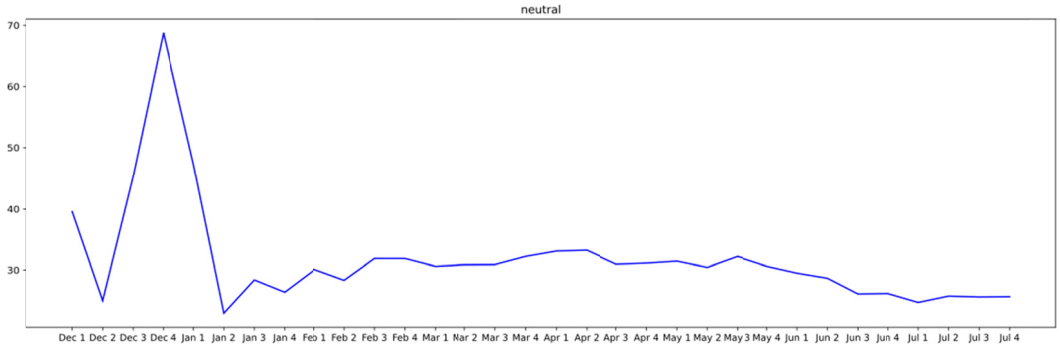

Figure 11. Actual Function of Neutral Sentiment

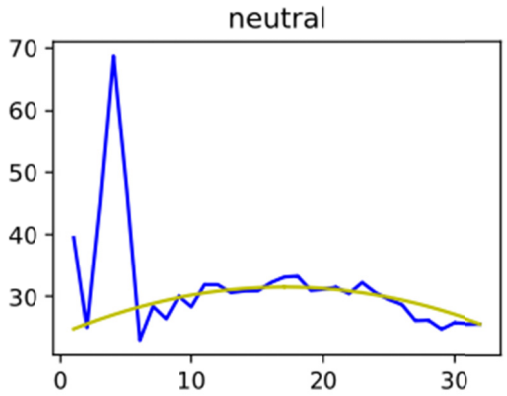

Figure 12. Actual vs Regression Function of Neutral Sentiment

Figure 11 represents the actual function, and the second is the comparison between the actual and the predicted. In the first graph, the x-axis is the period, written by the week's number in a month. The y-axis represents the percentage of that sentiment in all messages. In Figure 12, the x-axis is the periods of weeks from the first week of December, the same as the graph mentioned earlier. The y-axis is the percentage of the "neutral" sentiment, while the blue line represents the actual percentage, and the yellow line indicates the predicted proportion.

It can be seen from Figure 10 and Figure 12 that the pattern of the "neutral" sentiment proportion has some linearity. If we discard the unstable data in December and the beginning of January, we can generally see that the percentage of neutral increase at first until April and then decrease after that. It can be concluded that as the pandemic spreads, people from more countries start to report the coronavirus. Instead of passively receiving possibly unauthetic messages from some biased media people are now actually experiencing the disaster, hence might being more objective than merely conjecting the situation far away. The change in the "neutral" proportion can also be concluded as the supplement for the change in the percentage of positive and negative tags, which can be studied by further analysis of those four tags.

6.1.5 Discussion of the “Neutral towards Negative” Sentiment

By using regression model based on neural network, we examine the data of the “neutral towards negative” sentiment, excluding those of the first five periods of weeks which are fluctuating, and calculate the best-fit regression function, which can be expressed as $y = 98.4624x^2 - 96.8005x + 32.6666$, where x is the number of periods of weeks divided by 100, drawn as Figure 13:

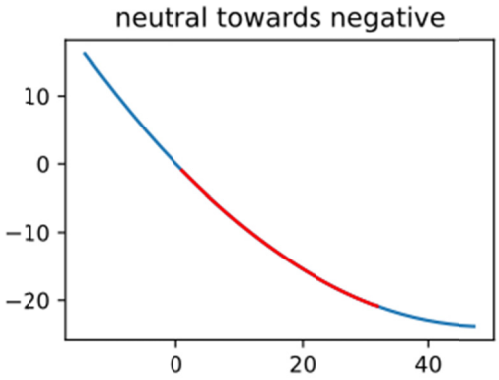

Figure 13. Regression Function of Neutral Towards Negative Sentiment

In Figure 13, the x-axis is the period number from the first week of December, the same as the graph mentioned earlier. The y-axis is the predicted percentage of the "neutral towards negative" sentiment with a unit of 1%. The red part represents the prediction of the function in the definition domain of the December to July data, and the blue part is for all valid domain of definition.

The graph for the actual function and comparison between it and the predicted function is as Figure 14 and Figure 15:

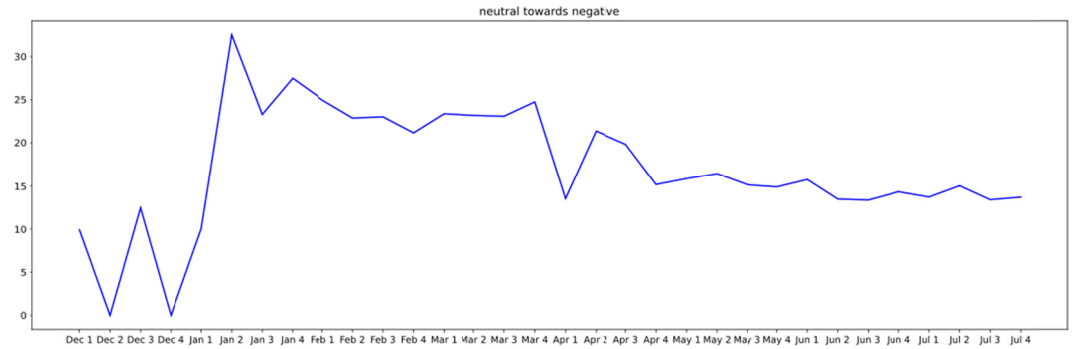

Figure 14. Actual Function of Neutral Towards Negative Sentiment

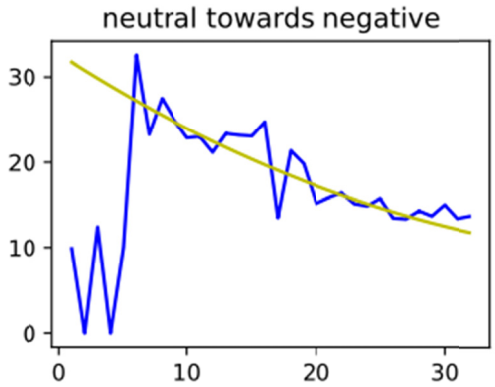

Figure 15. Actual vs Regression Function of Neutral Towards Negative Sentiment

Figure 14 represents the actual function, and the second is the comparison between the actual and the predicted. In the first graph, the x-axis is the period, written by the week's number in a month. The y-axis represents the percentage of that sentiment in all messages. In Figure 15, the x-axis is the periods of weeks from the first week of December, the same as the graph mentioned earlier. The y-axis is the percentage of the "neutral towards negative" sentiment, while the blue line represents the actual percentage, and the yellow line indicates the predicted proportion.

It can be seen in the Figure 15 that, besides the unstable data in December, the percentage of the "neutral towards

negative" sentiment generally shows a decreasing trend. This is corresponding to the patterns of those two positive tags. It can be conjectured that people express fewer negative messages as they are accustomed to their new lives and have more hope for the end of the pandemic.

6.1.6 Discussion of the "Negative" Sentiment

By using regression model based on neural network, we examine the data of the "negative" sentiment, excluding those of the first five periods of weeks which are fluctuating, and calculate the best-fit regression function, which can be expressed as $y = -22.1312x^2 + 4.5248x + 1.3617$, where x is the number of periods of weeks divided by 100, drawn as Figure 16:

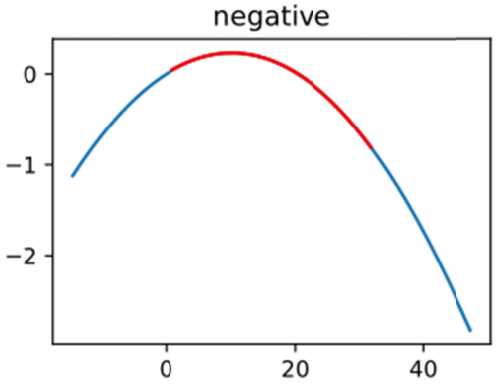

Figure 16. Regression Function of Negative Sentiment

In Figure 16, the x-axis is the period number from the first week of December, the same as the graph mentioned earlier. The y-axis is the predicted percentage of the "negative" sentiment with a unit of 1%. The red part represents the prediction of the function in the definition domain of the December to July data, and the blue part is for all valid domain of definition.

The graph for the actual function and comparison between it and the predicted function is as Figure 17 and Figure 18:

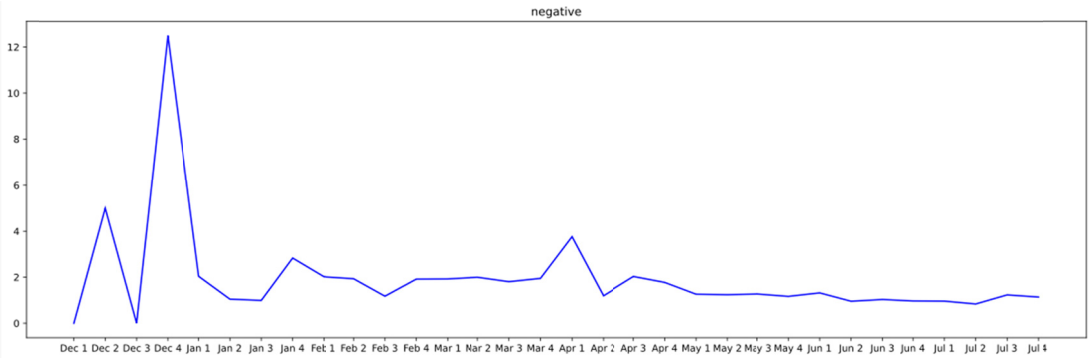

Figure 17. Actual Function of Negative Sentiment

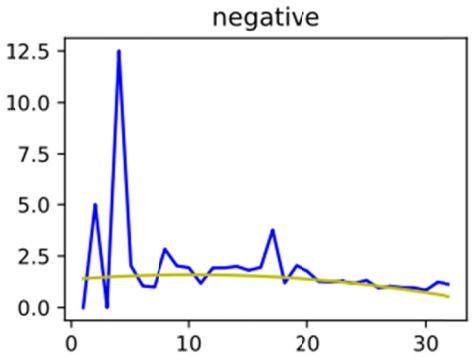

Figure 18. Actual vs Regression Function of Negative Sentiment

Figure 17 represents the actual function, and the second is the comparison between the actual and the predicted. In the first graph, the x-axis is the period, written by the week's number in a month. The y-axis represents the percentage of that sentiment in all messages. In Figure 18, the x-axis is the periods of weeks from the first week of December, the same as the graph mentioned earlier. The y-axis is the percentage of the "negative" sentiment, while the blue line represents the actual percentage, and the yellow line indicates the predicted proportion.

As is predicted, it can be seen in the graph that, besides the fluctuating data in December, the proportion of the "negative" sentiment has a decreasing trend. As is discussed, this trend might be due to the accommodation of a different life and the bettering future. Of course, since the proportion of the negative messages is so small, there might be no conclusion at all that can be drawn.

6.2 Analysis of the Typical Data Based on Gephi

6.2.1 Construction of the Gephi Graph Based on Force Atlas

We import the final data into Gephi and use Force Atlas, a Force-directed graph drawing algorithm mimicking the physical rules in reality, to construct a graph, which is shown as Figure 19:

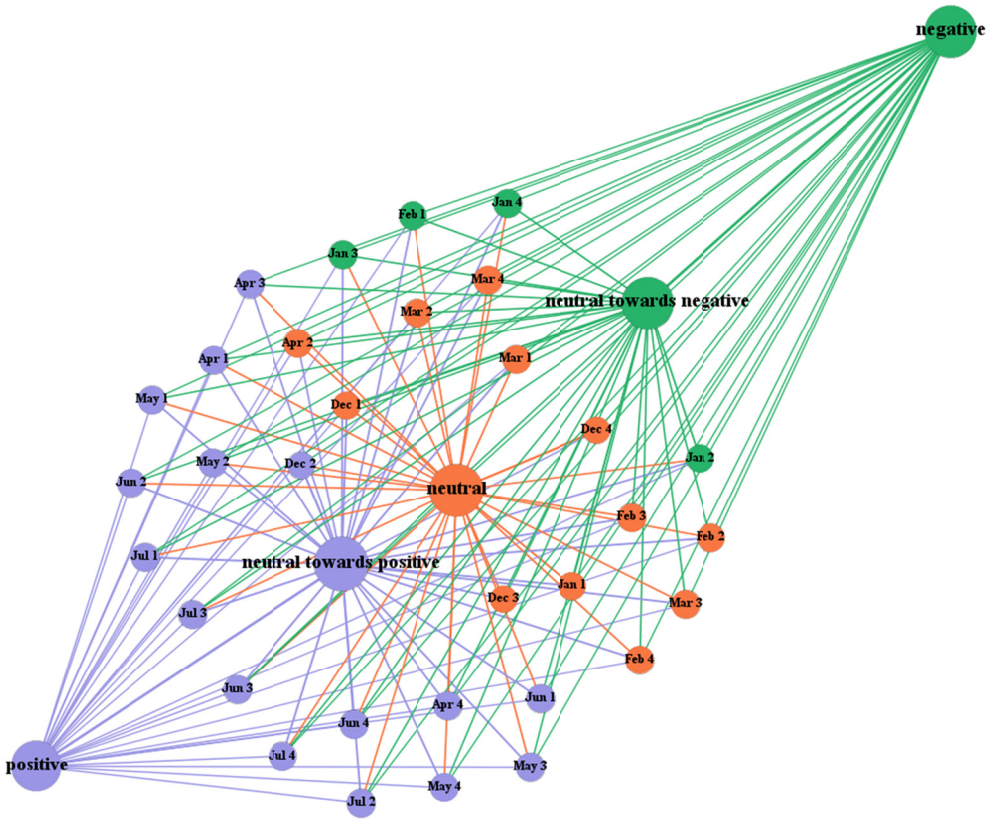

Figure 19. Gephi Graph of the Period's Relationship

In Figure 19, the nodes are the five sentiment tags and the periods. A tag node and a time period node are connected if the proportion of the tag's sentiment isn't zero in that week. The weight of the edge is the proportion of the related tag's sentiment in the related time period with a unit of 1%, that is to say, a period node with 23% messages being "neutral" will have an edge with a weight of 23 with the node "neutral." There is a 100-weighted edge between "positive" and "neutral towards positive" and between "negative" and "neutral towards negative." The Force Atlas algorithm used to draw the graph indicates that the period's property correlates with the relative spatial position. Two nodes fairly close to each other, or symmetrical about the centerline formed by the five sentiment tags, have similar traits. Meanwhile, nodes far away from each other most likely have significantly different properties.

In Figure 19, the relative sizes of the nodes represent the magnitude of their PageRank indexes. The color of the nodes indicates their modularity class calculated with a resolution index of 0.6. The relative distance between the tag nodes and the modularity class shows the relative sentiment polarity of the messages in that period. It can be seen from the graph that, besides December, the latter the period is, more positive the messages at that time are.

The result is corresponding to the Figure 20, in which the x-axis represents the period and the y-axis indicates the type of the sentiment tag.

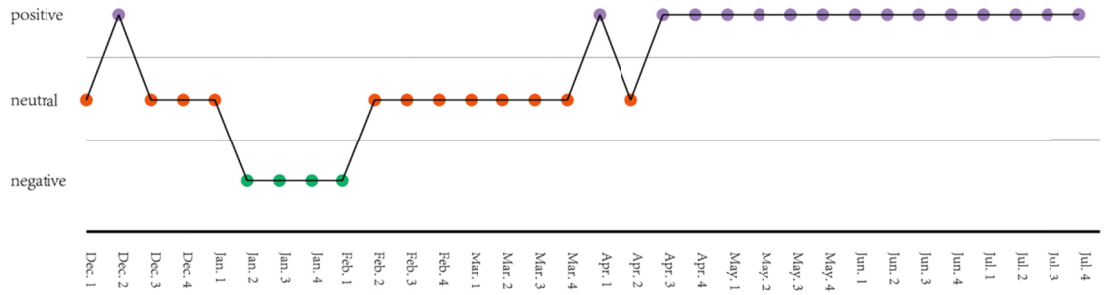

Figure 20. General Sentiment Tag by Time Period

6.2.2 Discussion of the Typical Data Based on Gephi

According to Figure 19, significant position change occurs between the nodes of the 1st week and the 2nd week in January, two adjacent periods. Figure 3 also demonstrates a considerable increase in negative proportion and indicates its pessimistic tendency, corresponding to Figure 19. This change is probably due to the events at the beginning of January. On January 7th, the end of the first week, scientists confirmed the discovery of COVID-19, the novel coronavirus. (Note 4) And on January 13th, the coronavirus spread to Thailand, the first case outside China. (Note 5) The situation reminded the people of the SARS in 2003. Users began to worry about the coming pandemic, sending messages with negative sentiment because of their anxiety about the future and, to some degree, the inefficient actions of the Chinese government, factors which contributed to the significant differences between the spatial position of the two nodes.

In Figure 19, there is also a remarkable difference between the positions of the 3rd period and the 4th period in January. On January 23rd, Wuhan, a city in China, started its total lockdown, prohibiting all transportation, including cars, trains, and flights. Additionally, the United States government declared the COVID-19 a public health emergency. Given the significant outbreak around the world, there were naturally many pessimistic messages and condemnations of China, leading to the difference between the two time periods.

Another typical example of the difference in Figure 19 is the contrast between the 3rd period and the 4th period in April. In Figure 3, the percentage of positive messages increases, and that of negative ones decreases. This trend is probably due to the "One World: Together at Home Global Concert" supported by the WHO on April 18th, (Note 6) which cheered the people up. As a hopeful and joyful sense starts to spread, naturally there will be more positive messages on the Internet.

7. Conclusion and Prospect

This article mainly puts forward the method of analyzing the sentiment proportion trend during a period by using the sentiment analysis model based on the neural network of Long Short-Term Memory, regarding the influence of the coronavirus pandemic on public's sentiments in messages on social media, evaluating the change of the proportion of different sentiments. The main conclusion of the article is as follows:

(1) Implement the LSTM architecture based on RNN to train a sentiment analysis neural network model using the sample data of movie reviews with high model efficiency.

(2) Apply the sentiment analysis model to calculate the sentiment score, analyze the sentiment proportion trend during the coronavirus period, and relate the pattern to significant events.

Of course, the article's method has several demerits, including the different sources of the two sentiment score data and the relatively small sample size. Further studies are needed if the trend shall be understood, and a more profound conclusion is to be drawn.

References

Da L., Rafal R., Michal P., & Kenji A. (2020). HEMOS: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media. *Information Processing and Management, 57*(6), 102290. https://doi.org/10.1016/j.ipm.2020.102290

Usman N., Imran R., Katarzyna M., & Muhammad I. (2020).Transformer based deep intelligent contextual embedding for Twitter sentiment analysis. *Future Generation Computer Systems, 113*, 58-69. https://doi.org/10.1016/j.future.2020.06.050

Notes

Note 1. WHO Coronavirus Disease (COVID-19) Dashboard. WHO. Retrieved 25 August 2020. https://covid19.who.int

Note 2. Ran, Geva. (2020). free dataset from news/message boards/blogs about CoronaVirus (4 month of data -5.2M posts). IEEE Dataport. http://doi.org/10.21227/kc4v-q323

Note 3. Rabindra, Lamsal. (2020). Coronavirus (COVID-19) Geo-tagged Tweets Dataset. IEEE Dataport. http://doi.org/10.21277/fpsb-jz61

Note 4. Khan, Natasha. (9 January 2020). New Virus Discovered by Chinese Scientists Investigating Pneumonia Outbreak. The Wall Street Journal. Retrieved 8 February 2020. https://www.wsj.com/articles/new-virus-discovered-by-chinese-scientists-investigating-pneumonia-outbreak-11578485668

Note 5. WHO | Novel Coronavirus – Thailand (ex-China). WHO. 14 January 2020. Retrieved 15 January 2020. https://www.who.int/csr/don/14-january-2020-novel-coronavirus-thailand-ex-china/en/

Note 6. Clements, Laura. (March 20, 2020). The best live streamed gigs to watch at home during coronavirus. walesonline.
https://www.walesonline.co.uk/whats-on/music-nightlife-news/live-streamed-gigs-coronavirus-yungblud-17955670