Jurnal Kejuruteraan SI 5(2) 2022: 177-189
https://doi.org/10.17576/jkukm-2022-si5(2)-19

Emotion Recognition and Analysis Of Netizens Based On Micro-Blog During Covid-19 Epidemic

(Pengiktirafan Emosi Dan Analisis Netizen Berdasarkan Micro-Blog Semasa Epidemik Covid-19)

Jiao BianBian[a], R. Leelavathi[a*], N. Lohgheswary[b] & Z. M. Nopiah[c]

[a]*Centre for Software Engineering, Faculty of Engineering, Built Environment & Information Technology, SEGi University, Kota Damansara, 47810 Selangor Darul Ehsan, Malaysia,*
[b]*Department of Electrical and Electronics Engineering, Xiamen University Malaysia, Sepang, 43900 Selangor Darul Ehsan,*
[c]*Department of Engineering Education, Faculty of Engineering & Built Environment, University Kebangsaan Malaysia*

**Corresponding author: leelavathiraj@segi.edu.my*

Received 15th June 2022, Received in revised form 31stJuly 2022
Accepted 1th September 2022 , Available online 15th November 20228

ABSTRACT

The research is about emotion recognition and analysis based on Micro-blog short text. Emotion recognition is an important field of text classification in Natural Language Processing. The data of this research comes from Micro-blog 100K record related to COVID-19 theme collected by Data fountain platform, the data are manually labeled, and the emotional tendencies of the text are negative, positive and neutral. The empirical part adopts dictionary emotion recognition method and machine learning emotion recognition respectively. The algorithms used include support vector machine and naive Bayes based on TFIDF, support vector machine and LSTM based on wod2vec. The five results are compared. Combined with statistical analysis methods, the emotions of netizens in the early stage of the epidemic are analyzed for public opinion. This research uses machine learning algorithm combined with statistical analysis to analyze current events in real time. It will be of great significance for the introduction and implementation of national policies.

Keywords: Micro-blog; emotion recognition; COVID-19; natural language processing

ABSTRAK

Kajian ini adalah berkenaan pengiktirafan dan analisis emosi berdasarkan teks singkat Mikro blog. Pengiktirafan emosi ialah bidand penting dalam klasifikasi teks dalam Pemprosesan Bahasa Semulajadi. Data penyelidikan ini didapi daripada rekod Mikro-blog 100K yang berkaitan dengan tema COVID-19 yang dikumpul oleh platform Datafountain, data tersebut dilabel secara manual dan kecenderungan emosi teks adalah negative, positif dan neutral. Bahagian empirikal menggunakan kaedah pengiktirafan emosi kamus dan pengiktirafan emosi pembelajaran mesin. Algoritma yang digunakan termasuklah sokongan mesin vector dan naif Bayes berdasarkan TFIDF, sokongan mesin vector dan LSTM berdasarkan wod2vec. Lima keputusan dibandingkan. Emosi netizen di peringkat awal epidemic dianalisis dengan menggabungkan kaedah analisis statistik untuk pendapat umum. Kajian ini menggunakan algoritma pembelajaran mesin yang digabungkan dengan analisis statistik untuk menganalisis peristiwa semasa dalam masa nyata. Ini sangat penting untuk pengenalan dan pelaksanaan dasar negara.

Kata kunci: Mikro blog; pengiktirafan emosi; COVID-19; pemprosesan bahasa semulajadi

INTRODUCTION

New Coronavirus pneumonia outbroke in early November 2019 in Wuhan, Hubei, China, Chinese prevention and control headquarters had taken effective measures, other provinces have provided point-to-point support to Hubei, and people in various regions of the country have actively cooperated (Zakaria and Singh 2021). At present, there has been an inflection point in China's domestic epidemic, but the international situation is not optimistic. By now, the virus has spread worldwide.

During the whole epidemic period, many people all over the world had to carry out various activities at home. Enterprises actively organized online office and schools organized online teaching. These changes and epidemic trends have different repercussions in people's psychology.

Some people are anxious about the epidemic and not used to working online, Others are happy for this sudden long holiday, because they work at home with their parents and children. Some people are worry about their health and afraid of being infected by the virus. In response to different emotions of different people, China has carried out a series of public opinion surveys and public opinion analysis. As the second largest social media in China, Micro-blog had 316 million mobile phone users by the first half of 2018, accounting for 32.6% of all mobile Internet users. Netizens often express their views and feelings on politics, entertainment, society and other issues on social media, and express some tendentious views on current affairs and policies. For the coronavirus emergency, this research collected the Micro-blog data during the rapid growth of the epidemic from January 1, 2020 to February 20, 2020 and analyzed the emotions of netizens during the epidemic.

Based on the novel coronavirus pneumonia related micro-blog news and its comments, this research analyzes the impact of the COVID-19 during the epidemic on Netizens' emotions based on the establishment of the epidemic situation and the emotional portrait of netizens.

Recently the economy of China is developing rapidly, and the society is full of vitality, but at the same time, there are many risks and contradictions. For example, the coronavirus has led to a great slowdown in China's macroeconomic growth. Similar events or policies will have a large or small impact on people's life and produce some different public opinion emotions. The masses' emotions are easily misled by organizations and individuals, resulting in bad masses' antisocial behavior. Therefore, the state must find these scattered and spontaneous contradictions in time and break them one by one as soon as possible to avoid them from growing into serious organized group confrontation or even large-scale fierce conflict. How to avoid such things and how to really understand the psychology and behavior of the masses, we need to analyze the public opinion of the mass media that actively share the emotions.

With the vigorous development of Internet related technology, the amount of public opinion information explosion has increased. One is the huge amount of data, the second is a variety of categories, and the third is the complicated information background. Such emergencies as New Coronavirus and other real-time hot events, it will immediately trigger strong reactions from various social groups, individuals and different political forces. In this case, the use of advanced computer technology, statistical technology and relevant algorithms for extraction, processing, analysis and judgment can make a rapid and comprehensive social problem early warning mechanism for event public opinion, and provide strong support for governments at all levels to accurately grasp the situation of network public opinion scientifically and efficiently do a good job in prevention and control publicity and public opinion guidance.

This research mainly analyzes the emotion of Micro-blog short texts in social media during the outbreak of COVID-19. Based on this, the government needs to analyze the public emotion tendency and master the public emotion state in time. This research divides the emotion state into negative, neutral and positive, monitors the public opinion information, puts forward corresponding policies, or downgrades or upgrades the existing policies to control the epidemic and keep people's material and spiritual normality to the greatest extent.

This research focuses on the key issues in emotion recognition and emotion recognition toward the Micro-blog content posted during COVID-19 epidemic, specifically, during January 1, 2020 to February 20, 2020, which is a period of rapid development of the epidemic. The emotional state of people during this period can basically reflect the emotional state of people during the whole epidemic period. This includes comparing the emotion recognition result on five models and analyzing the emotional state of micro-blog content during COVID-19 epidemic and the effect of COVID-19 on Netizens' emotions.

At present, the popular emotion recognition methods in academic circles have two branches, one is the method based on emotion dictionary. After constructing the emotion dictionary, match and count all kinds of words in the data set and choose different weights and operation methods for different types of words, to calculate the emotion score and further obtain the emotion polarity. The other method is based on machine learning algorithm. Firstly, the corpus is labeled, then different feature processing is carried out according to the characteristics of different languages, and then the algorithm is used to classify the data set. Text emotion recognition methods have developed greatly, especially English text analysis methods. Based on the real-time information related to the twitter election, Cambridge Analytics has established an emotion recognition system and recommendation system to respond to the emotions of Internet users in time, to help trump win the election. Text analysis based on Chinese Micro-blog and English Twitter is quite different. Because of China's complex national conditions, large population, long history of Chinese culture and profound cultural heritage, Chinese Micro-blog not only has changeable themes, more messy contents and different lengths, but also has great differences in writing habits and expression methods in different regions, classes and ages. This study mainly compares the emotion recognition effects of several mainstream models on Chinese data sets, constructs an emotion recognition evaluation system more suitable for the characteristics of Chinese Micro-blog, and analyzes the impact of the epidemic on the emotion of Internet users on the basis of emotion recognition data.

BASIC CONCEPTS OF TEXT EMOTION RECOGNITION

Text emotion recognition is the process of processing and analyzing text data, emotion mining and taking corresponding measures using the mining results. Due to the continuous development of 5G network, hardware storage, cloud computing and other technologies, there are a large

number of user led comment data on the Internet (such as Taobao, Micro-blog in China) that urgently need to generate value for products, events and people. These review data express people's different emotional colors and tendencies towards different things or the same things at different time nodes.

Huffaker (2010) stated that text emotion usually can be divided into two categories: negative and positive. These two emotional polarities can be subdivided into joy, anger, sadness, happy and praise, neutrality and devaluation. once other users review these subjective comments, they can understand others' views on a commodity, service or event, so as to make their own final judgment or decision on the same or similar things.

MAIN METHODS INVOLVED IN EMOTION RECOGNITION

The research on emotion recognition began with Pang's (2002a) research on the classification of emotion tendency of film review based on supervised learning algorithm and Turney's (2002) research on the classification of emotion tendency of text based on unsupervised learning. Many emotion recognition studies will use the film review dataset. With the rapid development of data mining, many research results on emotion recognition continue to emerge and show different research topics and development trends. According to the sample granularity, it can be divided into fine-grained emotion recognition based on short sentences and text fragments and coarse-grained emotion recognition based on articles and paragraphs. Coarse grained emotion recognition is based on the overall emotional tendency of the object to a product or event. It mainly uses some method to extract the features and train the extracted feature vectors to obtain the overall emotional tendency of the article or sentence. Fine grained emotion recognition is proposed by Hu and Liu (2004) to identify object views based on feature words, view words, adjacent words and their co-occurrence frequency.

According to the experimental methods, text emotion recognition is divided into three methods: method based on emotional dictionary, method based on machine learning and method based on deep learning.

The method based on emotional dictionary mainly classifies the text by identifying the emotional words in the dictionary. Based on integrating dependency syntactic features, Liu et al. (2017) Used K-means clustering algorithm to build an emotional dictionary for emotion recognition of online product reviews. Based on the traditional emotional dictionary, Yang (2014), Turney (2003) and Zhao Changyu (2020) respectively introduced the Late Dirichlet Allocation (LDA) model and Point Wise Mutual Information (PMI) algorithm to study the expansion of emotion dictionary and text emotion recognition. The method based on emotional dictionary has excellent performance in emotion recognition in specific fields, but this method rarely considers the context information of the text. Due to the continuous emergence of network new words, this method cannot update the emotional dictionary in time.

The method based on machine learning uses machine learning algorithms such as support vector machine, Naïve Bayes and maximum enterprise to analyze textual emotion. Pang et al. (2002.b) manually mark the emotional feature words in film reviews, compare and analyze the effect of SVM, NB, ME and other algorithms in emotional recognition in film reviews. The experimental results showed that SVM classification is the best. Wikarsa et al. (2015) Used NB algorithm to conduct fine-grained emotion recognition of text. The method based on machine learning is based on the context of text, which requires many manual annotations of corpus. This method is time-consuming and labor-consuming, and there are some problems such as inconsistent annotation.

At present, deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) have become a hot research direction in the field of text emotion recognition. Poria et al. (2015) Proposed a method to extract emotional features from short text based on deep CNN. This method uses the combined features of text, vision and audio to train the classifier based on multi-core learning. Compared with the existing methods, the emotion recognition accuracy of this method is improved by 14%. Wu Peng (2017) and Tu Manshu (2017) applied CNN to text emotion recognition, and their accuracy reached more than 93%.

In the research of Micro-blog emotion recognition, Zhang Haitao (2019) and others used Micro-blog data, based on complex network theory and based on the co-occurrence relationship between comment words, to build a sub-event network to dynamically track netizens' opinions and emotional fluctuations. Zhang Liu (2019) and Zeng Ziming (2019) respectively introduced multi-scale CNN and Bi-LSTM model to analyze the emotion of Micro-blog comment text. In the research of Micro-blog emotion recognition under the background of epidemic situation, An Lu (2017) constructed the emotional network map of epidemic situation stakeholders based on the forwarding relationship of Micro-blog users, and analyzed the emotional evolution trend of stakeholders according to public opinion topics. Zhou Honglei et al. (2020) built a topic emotion evolution model for the epidemic based on the situation awareness theory to explore the emotional changes of netizens behind the epidemic topic.

METHODOLOGY

This research is designed to analyze the emotion expressed by Micro-blog bloggers in China during COVID-19 epidemic period at a macro level. The quantitative method is appropriately adopted in this research. The data analysis will be performed in two parts: emotion recognition and emotion analysis.

The first part will perform emotion recognition by using 5 models. The results of it will include a summary data about precious, recall and F1 score for every model which

can be used to evaluate the applicability of the model and Micro-blog text emotion recognition. The second part will perform emotion analysis of netizens based on micro-blog during COVID-19.

DATA SOURCES

There are mainly two datasets namely micro-blog dataset and China's COVID-19 epidemic dataset. Micro-blog dataset is the primary dataset which was provided by DataFountain. Based on the COVID-19 related keywords, data collection was conducted. This dataset contained 1 million micro-blog data from January 1, 2020 to February 20, 2020, and 100 thousand data were manually tagged. The tagging was divided into three categories, 1 (positive), 0 (neutral) and -1 (negative). The dataset includes two training datasets: nCoV_100k_train.labled.csv, nCoV_900k_train.unlabled.csv and one test dataset: nCov_10k_test.csv. these three datasets are mainly used for emotion recognition. The dataset named nCoV_100k_train.labled.csv is also used for emotion analysis.

China's COVID-19 epidemic dataset is provided by the WHO. This dataset is the daily number of cases and deaths collected by WHO during January 3, 2020 and February 20, 2020. It can be used for emotion analysis.

EMOTION RECOGNITION OF NETIZENS BASED ON MICRO-BLOG

Firstly, the dataset named nCoV_100k_train.labled.csv are preprocess, after removing duplicates and deleting invalid values, there are 90024 records left. The remaining pieces of data are divided into 81021 records training set and 9003 records test set according to the ratio of 9:1, and each emotion category is also divided according to 9:1, so that the data of each category of the training set and the test set are balanced as much as possible and obey the same distribution, which ensure the trained model perform better.

The development platform of this experiment is Intel core i5-8250u CPU processor, the development tool is Python, and the corresponding version is 3.7.4; The deep learning library is keras, and the corresponding version is 2.3.1.

The experiment result will be analyzed, precision, Recall, F1 score and Macro average are calculated as evaluation indexes for evaluation of text emotion recognition method.

Description of evaluation indicators is:
Precision: the ratio of the number of real cases in the classification results obtained by the model to the number of positive cases obtained by the classifier, which can also be called the precision rate.
Recall: in the classification results obtained by the model, the ratio of the number of real cases to the number of positive cases in the sample can also be called recall.
F1 score: The harmonic average of accuracy and recall.
Macro average: The arithmetic average of all categories F1 score.

Symbol description is:

- M : number of categories.
- M_i : actual number of texts in Category i.
- $C_{r,i}$: number of correctly classified categories i.
- $C_{w,i}$: the quantity misclassified to category i by other categories

The evaluation indexes are calculated as follows:
The precision of category i is:

$$precision_i = \frac{C_{r,i}}{C_{r,i} + C_{w,i}}, \quad (1)$$

The Recall of category i is:

$$recall_i = \frac{C_{r,i}}{m_i} \quad (2)$$

The F1- Score of category i is:

$$F_i = \frac{2 \times recall_i \times precision_i}{recall_i + precision_i} \quad (3)$$

The precision of all samples is:

$$precision = \frac{\sum_i^m precision_i}{m} \quad (4)$$

The Recall of all samples is:

$$recall = \frac{\sum_i^m recall_i}{m} \quad (5)$$

The Macro average of all samples is:

$$MA = \frac{\sum_i^m F_i}{m} \quad (6)$$

EMOTION ANALYSIS OF NETIZENS BASED ON MICRO-BLOG

The dataset named nCoV_100k_train.labled.csv and China's COVID-19 epidemic dataset are used for emotion analysis. The data analysis will be performed in two phases.

In the first phrase, we mainly perform descriptive analysis on the dataset named nCoV_100k_train.labled.csv. Descriptive analysis is used to describe the overall situation of quantitative data. The concentration and difference characteristics of data are calculated through descriptive analysis to understand the basic situation of data. Therefore, descriptive analysis is often carried out first, and then in-depth analysis is carried out on the basis of it.

In descriptive analysis, common descriptive indicators include mean, median, mode, quartiles, standard deviation, standard error of the mean, frequencies, proportions and so on.

In second phrase, we mainly perform secondary analysis on the dataset named nCoV_100k_train.labled.csv and China COVID-19 epidemic dataset. Correlation is used to summarize the relationship between the emotion and COVID-19. Pearson correlation coefficient is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}. \quad (7)$$

In this research, the data process and analysis both use Python. Python is one of the most important languages in the computer world and a mainstream data analysis language. At present, the mainstream data analysis languages are python, R and MATLAB. Among them, python has rich and powerful libraries. It is often called glue language. It can easily connect various modules made by other languages. It is a more easy and rigorous programming language. The commonly used class libraries for Python data analysis include IPython, numpy, SciPy, pandas, Matplotlib, scikit-learn and Spyder.

Insertion Data

Firstly, we open nCoV_100k_train.labled.csv and WHO-COVID-19-china-data.csv using Python and perform data preview. Figure 1 shows the top 5 records in nCoV_100k_train.labled.csv.

	微博id	微博发布时间	发布人账号	微博中文内容	微博图片	微博视频	情感倾向
0	4456072029125500	01月01日 23:50	存曦1988	写在年末冬初孩子流感的第五天，我们仍然没有忘记热情拥抱这2020年的第一天。带着一丝迷信，早...	['https://ww2.sinaimg.cn/orj360/005VnA1zly1gah...	[]	0
1	4456074167480980	01月01日 23:58	LunaKrys	开年大模型...累到以为自己发烧了腰疼膝盖疼腿疼胳膊疼脖子疼#Luna的Krystallife#?	[]	[]	-1
2	4456054253264520	01月01日 22:39	小王爷学辩论o_O	□偷癞空饩褪俏业□，爹，发烧快好，毕竟美好的假期拿来养病不太好，假期还是要好好享受快乐，爹，...	['https://ww2.sinaimg.cn/thumb150/006ymYXKgy1g...	[]	1
3	4456061509126470	01月01日 23:08	芩鎵	新年的第一天感冒又发烧的也太衰了但是我要想着明天一定会好的?	['https://ww2.sinaimg.cn/orj360/005FL9LZgy1gah...	[]	1
4	4455979322528190	01月01日 17:42	changlwj	问：我们意念里有坏的想法了，天神就会给记下来，那如果有好的想法也会被记下来吗？答：那当然了。...	[]	[]	1

FIGURE 1. The top 5 records in nCoV_100k_train.labled.csv.

As records show, the Micro-blog data format is shown in Table 1.

TABLE 1. The data format of Micro-blog dataset

Field name	Type	Description
Micro-blog ID	Integer	Publisher ID of microblog information
Micro-blog release time	Date	Date of releasing
Publisher account	string	Publisher ID of microblog information
Micro-blog Chinese content	Object	The detailed publish Chinese content
Micro-blog picture	Object	Publish picture is URL hyperlink, [] means no picture
Micro-blog video	Object	Publish video is URL hyperlink, [] means no video
Emotion tendency	Integer	The emotion tendency of Publisher, the value is {1, 0, -1}, 1 stands for positive, 0 stands for neutral, -1 stands for negative.

Figure 2 shows the top 5 records in WHO-COVID-19-china-data.csv.

	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
0	2020/1/3	CN	China	WPRO	0	0	0	0
1	2020/1/4	CN	China	WPRO	1	1	0	0
2	2020/1/5	CN	China	WPRO	0	1	0	0
3	2020/1/6	CN	China	WPRO	3	4	0	0
4	2020/1/7	CN	China	WPRO	0	4	0	0

FIGURE 2. The top 5 records in WHO-COVID-19-china-data.csv.

As records show, the Chinese COVID-19 epidemic data format is shown in Table 2.

TABLE 2. The data format of China's COVID-19 epidemic dataset

Field	Type	Description
Date_reported	Date	Date of reporting
Country_code	String	ISO Alpha-2 country code
Country	String	Country, territory, area
WHO_region	String	WHO regional offices
New_cases	Integer	New confirmed cases. Calculated by subtracting previous cumulative case count from current cumulative cases count.*
Cumulative_cases	Integer	Cumulative confirmed cases reported to WHO to date.
New_deaths	Integer	New confirmed deaths. Calculated by subtracting previous cumulative deaths from current cumulative deaths.*
Cumulative_deaths	Integer	Cumulative confirmed deaths reported to WHO to date.

Data Analysis and Discussion

In data analysis phrase, we need import several libraries to perform data processing, numerical calculation and data analysis. These libraries include NumPy, SciPy, Matplotlib and pandas. Among them, NumPy is used for data scientific calculation, SciPy is used for hypothesis test, Matplotlib is used for data visualization, pandas are used for data preprocessing and data statistical analysis which is very important in this research.

Because of the limitation of computing power, this research adopts the dataset during January 3, 2020 and February 20, which includes 1 million Micro-blog records.

In this study, it is first assumed that the emotion is affected by COVID-19, so COVID-19 data is the independent variable, and the emotion is the dependent variable.

Data Preprocessing

Before data analysis, the data we get may have missing values, duplicate values, etc., which need data preprocessing before use. Data preprocessing generally includes missing value processing, feature normalization, discretization and continuity, denoising.

As "Emotion tendency" is a key variable in data analysis, we process this variable firstly. Figure 3 shows the basic characteristics of "Emotion tendency".

```
0      57619
1      25392
-1     16902
4          1
9          1
-          1
.          1
10         1
-2         1
Name: 情感倾向, dtype: int64
```

FIGURE 3. The basic characteristics of "Emotion tendency"

As shown in Figure 3, there are only 99919 emotion tendency values, additionally, Table 1 indicated that the value of "emotion tendency" should only be -1, 0 and 1, so there are only 99913 valid values excluding 81 missing values and 6 invalid values.

Figure 4 shows other missing values in the Micro-blog dataset. All the missing values are deleted following the listwise delete strategy.

```
微博id          0
微博发布时间        0
发布人账号         0
微博中文内容      353
微博图片          0
微博视频          0
情感倾向          0
dtype: int64
```

FIGURE 4. The missing values in the Micro-blog dataset

In term of China's COVID-19 epidemic dataset, the first record is January 3, 2020 which is different from start date January 1, 2020 in Micro-blog dataset. In order to maintain time consistency, we supplemented the data of the previous two days. In January 1, 2020 and January 2, 2020, all the number of cases is 0.

Descriptive Analysis

Descriptive analysis is used to describe the overall situation of quantitative data. The concentration and difference characteristics of data are calculated through descriptive analysis to understand the basic situation of data. Therefore, descriptive analysis is often carried out first, and then in-depth analysis is carried out on the basis of it.

The frequency analysis is performed on "Emotion tendency", the analysis result is shown in Figure 5. As shown in the Figure 5, the number of neutral emotion accounts for more than half, followed by positive emotion and negative emotion.

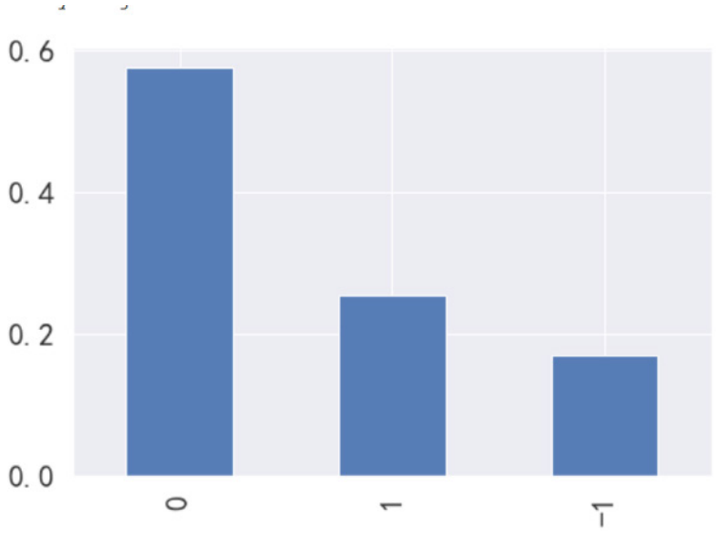

FIGURE 5. "Emotion tendency" frequency proportion histogram

In term of the length of Micro-blog comment, the frequency analysis is shown in Figure 6. Most of the comments are distributed in about 150 words, the number of less than 150 words is basically maintained at 2000, and there are few microblog contents longer than 150 words. Therefore, when the sentence length is set by the neural network, it can be set to a value of 200 or higher than 150 to avoid losing too much information.

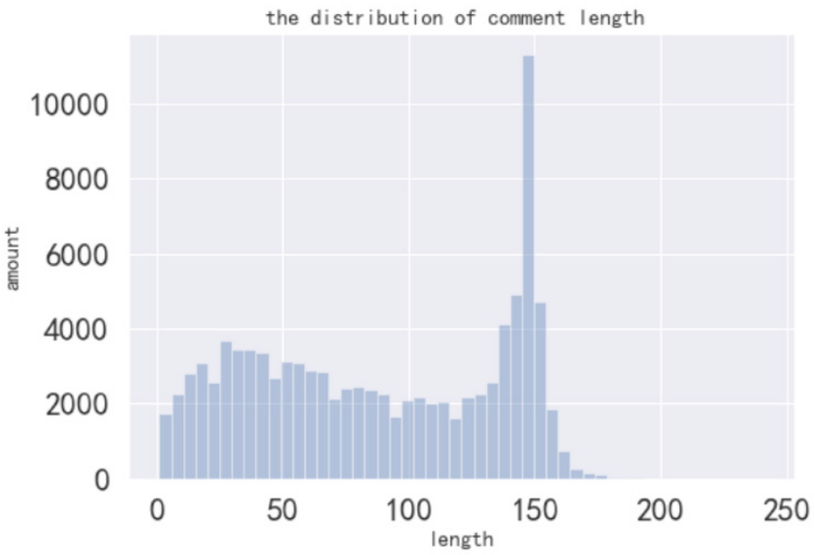

FIGURE 6. The distribution of comment length.

Previous descriptive analysis is focused on the external attributes of emotion tendency and the text length, proportion, etc. furthermore, we analyze the content of the text. The theme of the text can be observed through word frequency. Figure 7 is the word cloud chart.

FIGURE 7. Word cloud chart for Micro-blog content

As shown in Figure 7, we can see that the content mainly focuses on the topic of COVID-19, including some content of “Go China!” and “Go Wuhan!”, most of which are positive or neutral.

Correlation

To better understand the change of emotion tendency during January 1, 2020 and February 20, we analyze the number of positive, negative and neutral emotion based on date. Figure 8 shows the distribution map of microblog number respectively, and Figure 9 shows the distribution map of emotion proportion of microblog. The red line indicates the number of neutral emotions, the green line indicates the number of positive emotions, the blue line indicates the number of negative emotions.

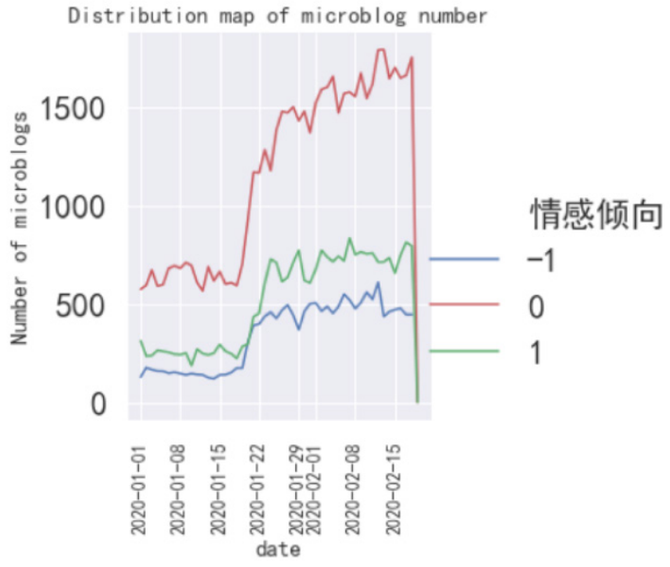

FIGURE 8. The distribution map of microblog number

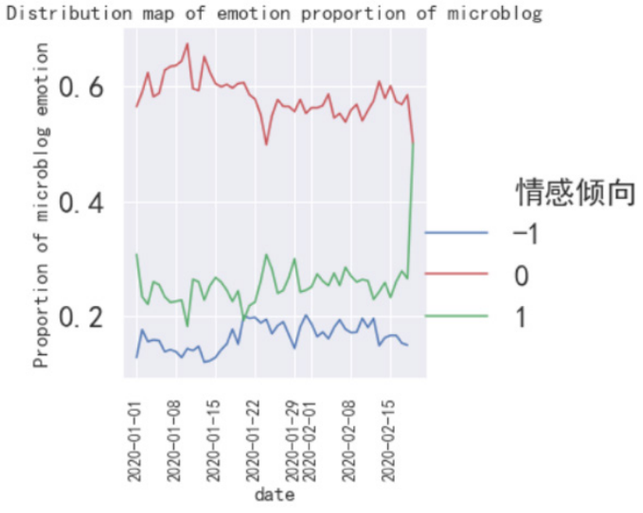

FIGURE 9. The distribution map of emotion proportion of microblog

As shown in Figure 8 and Figure 9, The netizens with neutral emotion accounted for 60%, and the netizens with positive emotion were more than those with negative emotion. When the epidemic began to appear on a small scale, the people did not know much about its severity, and most of them were positive. COVID-19 has been reported in Wuhan since January 15th. In January 18th, Zhong Nanshan went to Wuhan and announced that COVID-19 could be transmitted to others. The national alarm was heard, and the atmosphere of anxiety and tension was permeated in the society. As shown in figure 3.6, the number of microblogs had increased rapidly and keep high for a long time. On January 23, Wuhan was lockdown, which made people put down some concerns for the time being. Then, although there was a series of growth in the number of confirmed cases every day, all parts of the country actively and quickly took countermeasures, as well as medical staff working overtime to fight against death, the number of cured people also continued to rise, the people's mood was basically stable, people seriously obeyed the national arrangement, and waited quietly at home for the end of the catastrophe.

One of important topic of this research is to identify the Correlation between netizens' emotions and COVID-19, specifically, whether netizens' emotions are influenced by the epidemic and the degree of influence.

Correlation is a mutual relationship between phenomenon, which describes the strength of a linear relationship between two variables. Correlation coefficient is the statistical index of the degree to which two variables are associated or related. A positive value indicates a positive correlation, a negative value indicates a negative correlation. The value of correlation coefficient indicates the intensity of correlation. There are many calculation methods for correlation coefficient, among which Pearson Product-moment correlation is the most common.

In this research, we adopt the Pearson Product-moment correlation to calculate the correlation between netizens' emotions and COVID-19. The independent variables include: "New_cases", "Cumulative_cases", "New_deaths", "Cumulative_deaths". The dependent variable cannot be obtained directly, so it needs secondary calculation and analysis. It is the number of emotions group by emotional polarity and date. After obtaining the dependent variable, the research calculates the correlation coefficient matrix which gives the correlation coefficient between any two variables. Figure 10 shows the correlation between COVID-19 and the number of positive emotions.

	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths	positive
New_cases	1.000000	0.586123	0.844387	0.535983	0.531506
Cumulative_cases	0.586123	1.000000	0.858729	0.996338	0.505203
New_deaths	0.844387	0.858729	1.000000	0.832142	0.593418
Cumulative_deaths	0.535983	0.996338	0.832142	1.000000	0.469435
positive	0.531506	0.505203	0.593418	0.469435	1.000000

FIGURE 10. the correlation coefficient matrix of positive emotion

As shown in Figure 10, the correlation between the number of positive emotions and New_cases, Cumulative_cases, New_deaths are all between 0.5 and 0.7, which indicates there is a medium relationship.

Figure 11 shows the correlation between COVID-19 and the number of negative emotions.

	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths	negative
New_cases	1.000000	0.586123	0.844387	0.535983	0.510699
Cumulative_cases	0.586123	1.000000	0.858729	0.996338	0.605217
New_deaths	0.844387	0.858729	1.000000	0.832142	0.609069
Cumulative_deaths	0.535983	0.996338	0.832142	1.000000	0.591061
negative	0.510699	0.605217	0.609069	0.591061	1.000000

FIGURE 11. the correlation coefficient matrix of negative emotion

As shown in Figure 11, the correlation between the number of negative emotions and New_cases, Cumulative_cases, ew_deaths, Cumulative_deaths are all between 0.5 and 0.7, which indicates there is a medium relationship. Compared to positive emotion, the correlation between negative emotion and COVID-19 is stronger.

Figure 12 shows the correlation between COVID-19 and the number of neutral emotions.

	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths	nutral
New_cases	1.000000	0.586123	0.844387	0.535983	0.580537
Cumulative_cases	0.586123	1.000000	0.858729	0.996338	0.532021
New_deaths	0.844387	0.858729	1.000000	0.832142	0.624768
Cumulative_deaths	0.535983	0.996338	0.832142	1.000000	0.491742
nutral	0.580537	0.532021	0.624768	0.491742	1.000000

FIGURE 12. The correlation coefficient matrix of neutral emotion

As shown in Figure 12, the correlation between the number of neutral emotions and New_cases, Cumulative_cases, New_deaths are all between 0.5 and 0.7, which indicates there is a medium relationship. Compared to positive emotion, the correlation between neutral emotion and COVID-19 is stronger.

Chi-square test

Chi square test is a widely used hypothesis test method for counting data. It belongs to nonparametric test, which mainly compares the correlation analysis of two or more sample rates (constituent ratio) and two classification variables. The fundamental idea is to compare the coincidence degree or goodness of fit between the theoretical frequency and the actual frequency.

In this research, emotion tendency is a classification variable, we perform the Chi-square test on emotion tendency and date. The cross table for emotion category and date is in Figure 13.

情感倾向	-1	0	1
date			
2020-01-01	130.0	574.0	313.0
2020-01-02	178.0	595.0	236.0
2020-01-03	168.0	673.0	238.0
2020-01-04	161.0	591.0	264.0
2020-01-05	160.0	598.0	259.0
2020-01-06	150.0	682.0	254.0
2020-01-07	155.0	694.0	245.0
2020-01-08	148.0	682.0	242.0
2020-01-09	141.0	710.0	251.0
2020-01-10	148.0	695.0	188.0

FIGURE 13. The cross-table for emotion and date

The H0 hypothesis: the emotion category and date are independent of each other.
The H1 hypothesis: the emotion category is related to date.

The significance level is set to 0.05, the calculated P value is greater than 0.05, so H0 holds, the emotion category and date are independent of each other. The experiment result is shown in Table 3.

TABLE 3. Experiment result

Model	Precious	Recall	MA
SNOWNLP	46.56%	42.13%	38.22%
TFIDF+SVM	58.35%	61.23%	59.11%
TFIDF+NB	60.00%	61.98%	51.34%
Word2vec+SVM	72.13%	72.35%	70.28%
Embedding+LSTM	82.60%	85.12%	83.34%

According to the experimental process and results, the following conclusions can be drawn. For the machine learning model, the uneven distribution of data in each category of the training set will greatly affect the performance of the final classifier. The model often performs well or even too well for samples with many categories, resulting in over fitting, while insufficient learning for samples with few categories leads to under fitting. For example, the recall rate of TFIDF+NB in neutral emotion recognition is 94.00%, while the recall rate of negative emotion is only 24%. The ratio of neutral emotion samples to negative emotion samples is close to 4:1.

Compared with the traditional text representation method of TFIDF, the word vector method has stronger representation ability of text, but its interpretation ability is relatively weak. This research is a classification of short text. In TFIDF representation, a text is limited to 20 nearly 30000-dimensional word vectors at most; The word vector method is obtained by averaging all texts with 200-dimensional word vectors. From the results in the table above, the algorithm using word vectors to represent texts is better than the traditional method in accuracy, recall and ma.

Compared with traditional machine learning algorithms, text classification using LSTM can achieve better classification results. Compared with support vector machine, LSTM improves the accuracy by nearly 10%, recall by 13% and Ma by nearly 13%. This research belongs to short text classification. Each sample can provide less and miscellaneous features. Compared with single-layer machine learning algorithm, high-dimensional neural network can extract more feature information from this kind of samples, so it performs better.

The algorithm based on affective dictionary performs well in similar texts, but the generalization ability of new text materials is not enough. In addition, the dictionary-based algorithm belongs to unsupervised learning, and multiple thresholds need to be formulated in multi classification, and the selection of this threshold depends on experience and subjective judgment. In this research, the score of emotion is less than 0.1 for negative emotion, greater than 0.1, less than 0.9 for neutral emotion, and greater than 0.9 for positive emotion. Because there are few feature words in short text, emotional words are not easy to capture, and the emotional dictionary used by SNOWNLP is different from the sample fields used in this research. In addition, the distribution of three types of samples in the data set is uneven. Finally, these two thresholds are selected to optimize each index.

CONCLUSION

Through analyzing the research datasets, the research found that netizens' emotion for the epidemic was neutral and positive in the early stage of the epidemic. Although people's emotions fluctuate to a certain extent with the development of the epidemic, which is shown in the sharp increase in the number of comments after the government disclosed that the epidemic can be transmitted from person to person, most people can rationally deal with the development of the epidemic and all kinds of emergencies.

According to the research on the netizens' emotion and epidemic factors by time period, there is a relatively strong correlation between COVID-19 epidemic factors and emotion tendency includes positive, negative and neutral from January 01, 2020 to February 20, 2020. However, the correlation has decreased significantly over time. This is inseparable from the positive actions of the government and the rapid response to Internet public opinion, which shows that our country has been able to respond to COVID-19's efforts effectively and has been recognized by netizens. At the same time, the release and supervision of epidemic information is also in place to avoid the influence of false and fake information on Netizens' emotions.

ACKNOWLEDGEMENT

The authors would like to thank SEGi University (SEGiIRF/2020-8/FoEBEIT-37/103), and Xiamen University Malaysia (XMURF/2022 C9/IECE/0026) and the Department of Engineering Education, Faculty of Engineering and Built Environment, UKM for supporting the research.

REFERENCES

An Lu, Ou Menghua. 2017. Study on social network emotion map of stakeholders in public health emergencies [J]. *Library and Information Work* 61 (20): 120-130

Hu M Q, Liu B. 2004. Mining opinion features in customer reviews [C]. Proceedings of the 19th National Conference on Artifical Intelligence. Palo Alto: AAAI Press: 755-760

Huffaker, D. 2010. Dimensions of leadership and social influence in online communities. *Human Communication Research* 36(4): 593–617.

Liu F, Wei F, Yu K. 2017. Sentiment classification of reviews on automobile website by combining word2vec and dependency parsing [C]. Proceedings of the International Conference on Smart Computing and Communication. Berlin, Germany: 206-221.

Pang B, Lee L, Vaithyanathan S. 2002. Thumbs up: Sentiment classification using machine learning techniques[C]. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. *Stroudsburg: Association for Computational Linguistics* 10: 79-86

Poria S, Cambria E, Gelbukh. 2015. Deep convolutional neural net-work textual features and multiple kernel learning for utterance level multimodal sentiment analysis [C]. Proceedings of the Conference on Empirical Methods in Natural Language Process- ing, Lisbon, Portugal: 2539-2544.

Tu, M., Zhang, Y. and Yan, Y. 2017. Emotion analysis of microblog events based on cnn-svm and forwarding tree [J]. *Information Engineering* 3 (3): 77-85.

Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics: 417-424

Turney, P. D., Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association [J]. *ACM Transactions on Information Systems* 21(4): 315-346.

Wikrsal, L., Thahir, S. N. 2015. A text mining application of emotion classifications of Twitter's users using Naïve Bayes method [C]. Proceedings of the 1st International Conference on Wireless and Telematics, Manado, Indonesia: 1-6.

Wu, P., Liu, H. and Shen, S. 2017. Research on online public opinion emotion recognition based on deep learning and OCC emotion rules [J]. *Journal of Information Technology* 36 (9): 972-980.

Yang, M., Zhu, D. J. and Choe, K. P. 2014. A topic model for building fine-grained domain-specific emotion lexicon [C]. Proceedings of 52nd Annual Meeting of the Association for Computational Lin-guistics, Baltimore, USA: 421-426.

Zakaria, S. A. S. and Singh, A. K. M. 2021. Impacts of Covid-19 outbreak on civil engineering activities in the Malaysian construction industry: A review. *Jurnal Kejuruteraan* 33(3): 477-485.

Zeng, Z. and Wan, P. 2019. Emotion analysis of public security event microblog based on double-layer attention and Bi LSTM [J]. *Information Science* 37(6): 23-29.

Zhang, H, Liu, Y. and Zhang, X. 2019. Research on topic discovery and emotional fluctuation of netizens based on modularity -- Taking the topic of "trade friction between China and the United States" on Sina Weibo as an example [J]. *Library and Information Work* 63(4): 6-14

Zhang, L., Wang, X. and Huang, B. 2019. Emotion classification model and experimental research of microblog comments based on multi-scale convolution neural network based on word vector [J]. *Library and Information Work* 63 (18): 99-108.

Zhao, C., Wu, Y. and Wang, J. 2020. "Belt and Road" initiative Twitter text topic mining and sentiment analysis [J]. *Library and Information Work* 63(19): 119-127.

Zhou, H., Zhang, H., and Zhang, X. 2020. Topic emotion map: the starting point of public opinion guidance for public health emergencies [J]. *Information Science* 38(7): 15-21.