World Scientific Research Journal Volume 9 Issue 4, 2023
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202304_9(4).0011

Sentiment Analysis of User Comments on COVID-19 Based on LSTM

Yi Yang[1, *]

[1]Faculty of science, The University of Sydney, Sydney, 2006, Australia

[*]yyan6506@uni.sydney.edu.au

Abstract

This paper mainly studies the emotional analysis of user comments on YouTube videos during the COVID-19 in the United Kingdom in the past year, so as to understand users' emotions and the topics they pay attention to, and assist relevant public opinion departments to quickly analyze public opinion. This paper uses LSTM as the emotion analysis model, trains and verifies the labeled data, and adjusts the parameters to select the better model. Besides, multi-level and multi-dimensional analysis have been conducted based on the emotional changes of users under the past epidemic prevention and control and LDA (Latent Dirichlet Allocation) subject extraction to help relevant public opinion departments better monitor online public opinion.

Keywords

COVID-19; LSTM; User sentiment; LDA.

1. INTRODUCTION

As an online video service provider in the current industry, YouTube's system processes tens of millions of video clips every day, providing high-quality video upload, distribution, display and browsing services for thousands of users around the world. It has also become the main platform for users to obtain various information. In recent years, the COVID-19 has raged, and users have made a lot of comments on YouTube. These comments with different concerns and emotions may lead to some public opinion, which will worsen the network ecology. Therefore, how to quickly collect users' comments, classify and analyze their emotional tendencies has become the focus of public opinion analysis.

Emotional analysis is an important task in natural language processing. It can process and analyze text with subjective emotion, and extract the corresponding emotional polarity to provide the basis for subsequent analysis. At present, the domestic public opinion sentiment analysis mainly analyzes the comment text from the perspective of time, space or text semantics [1], and applies the sentiment map to construct the situation of the public opinion in multiple periods to study the evolution direction of the emotion trend [2]. Foreign countries mainly use different emotion analysis methods to classify user emotions, and improve the model effect by optimizing the design of algorithms [3], and study the transmission mechanism and channels of public opinion in emergencies [4]. According to the research results at home and abroad in recent years, the existing research mainly focuses on model optimization and spatial-temporal domain analysis, but less on the different theme bias of the emotion of public opinion.

This paper mainly studies the following three aspects.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202304_9(4).0011

Compare various machine learning and deep learning models, select the best model, analyze the COVID-19 news comments in British under YouTube, and use the save training to predict the new COVID-19 news comments, so as to carry out further emotional analysis.

During the COVID-19 in the UK in the past year, the emotional changes of user comments on YouTube videos were extracted based on LDA, and the themes of positive and negative texts were analyzed in more detail.

This paper has certain practical value. It not only compares several models, but also selects the best model to conduct further emotional analysis of the text, and then conducts in-depth mining of the text based on time relations and LDA, providing a more representative network public opinion for public opinion supervision.

2. CONSTRUCTION OF USER COMMENT EMOTION MODEL BASED ON LSTM

2.1. Model construction

The LSTM (Long Short-Term Memory) model has three main mechanisms: forgetting gate, update gate and output gate [5]. The forgetting gate is used to filter irrelevant information, and the update gate and output gate are responsible for retaining useful information for a longer time. The specific structure inside the LSTM model unit is shown in Figure 1.

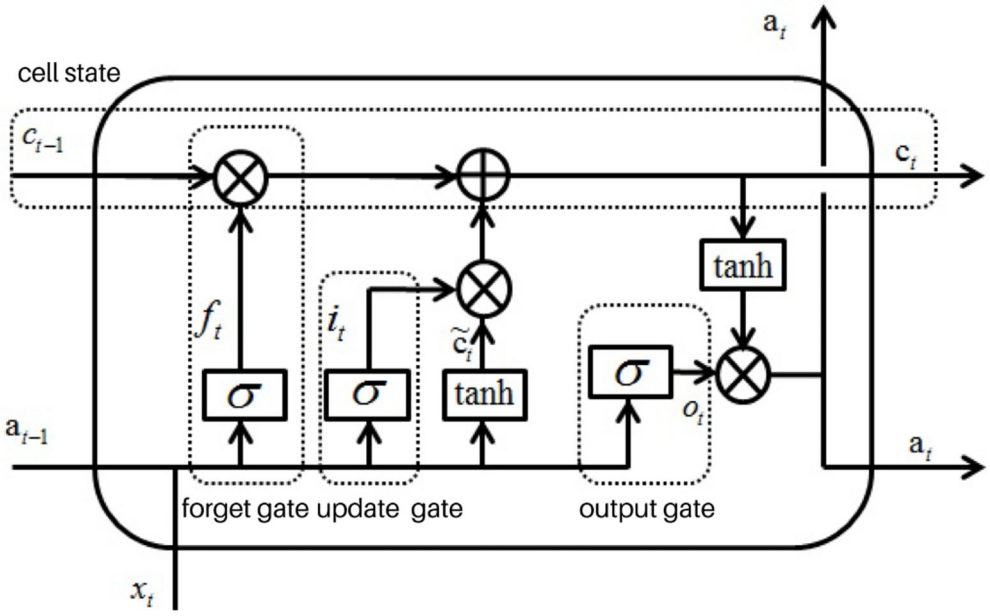

Figure 1. LSTM Model Unit Structure Diagram

LSTM controls long-term memory through cell state and short-term memory through hidden state, which improves the disadvantages of RNN and is more able to obtain long-distance dependent information than RNN. In the text data, the order of the text represents the time sequence. We use the word vector provided by the open source pre-training model word2vec as the "time sequence feature" of each word to input into the LSTM model to provide semantic text features.

2.2. Data preparation

We use selenium to simulate users' access to news, and collect as much user comment information as possible by visiting video pages, constantly pulling down and clicking on the user's response to each comment. Table 1 is the example of some comments.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202304_9(4).0011

Table 1. Examples of some comments

Comment
Boris did it just to calm the public so everyone forgot about his “partying” lol!
Be interesting to see if global actually airs this on Canadian news outlets
Good news to see the UK going in the right direction. Hopefully more countries will follow.

According to the above process, we first used snowNLP to mark the 1016 Canadian comments collected for the training and validation of the model, and collected the comment data on the British epidemic at the three date points of April 25, 2021, December 6, 2021 and January 20, 2022, which has 4605, 701 and 2512 respectively.

2.3. Labeling process

First of all, we annotate the collected Canadian review data. In order to reduce the labeling cost, we use the open source snowNLP to assist labeling. SnowNLP is a python toolkit with part-of-speech labeling and emotion analysis functions. It can provide emotion analysis functions based on thesaurus. We use it to perform preliminary labeling, and then conduct manual inspection to correct error tagging. Finally, we preprocess the labeled text data such as punctuation, number and expression removal。

2.4. Model construction and selection

After obtaining the labeled data, the effect of the model has been tested through the following process and the best model has been selected.

Randomly divide the data into training sets and test sets, with a ratio of 8:2.

Build different emotion classification models, train on the training set, and test on the test set, judge different models according to the results of the test set, and obtain the best model among them.

Adjust the parameters of the optimal model obtained in the previous step, and obtain the optimal parameters.

Train the model according to the optimal parameters and predict the new data.

This paper has compared SVM, logistic regression, Ridge, decision tree, random forest and LSTM models (the number of neurons is 64, the length of intercepted text is 18 on average), and evaluated the effect with accuracy and F1 value, the results are shown in Table 2.

Table 2. Prediction effect of each model

	SVM	Logistic Regression	Ridge	Decision Tree	Random Forest	LSTM
accuracy	0.6470588	0.6519608	0.6029412	0.6078431	0.6372549	0.669950739
F1	0.6392573	0.6482332	0.6028553	0.6075036	0.6369408	0.659233834

The above results show that the LSTM model is superior to other models in terms of accuracy and F1 value. In this paper, the LSTM model is selected for more in-depth experimental comparison, and the orthogonal parameter adjustment method is used to quickly and effectively find the better parameters.

We debug the number of neurons (hidden) and use the text length (seq) to train the LSTM model, compare the model effect with the accuracy and F1 of each parameter model as the evaluation index, and find that when seq is equal to 30, hidden is equal to 32, the LSTM model has the best effect, with the accuracy of 68.97% and F1 value of 69.36%. In the emotional

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202304_9(4).0011

analysis section, we will use the optimal model to predict and comment on emotions and conduct subsequent analysis.

3. EMOTIONAL ANALYSIS OF COVID-19 SITUATION

3.1. Analysis of user emotion change

Using the LSTM model constructed above, we analyzed the comments on the British COVID-19 on YouTube video at the three date points of April 25, 2021, December 6, 2021, and January 20, 2022 respectively, and marked positive and negative emotions and statistical proportions. The results are shown in Figure 2.

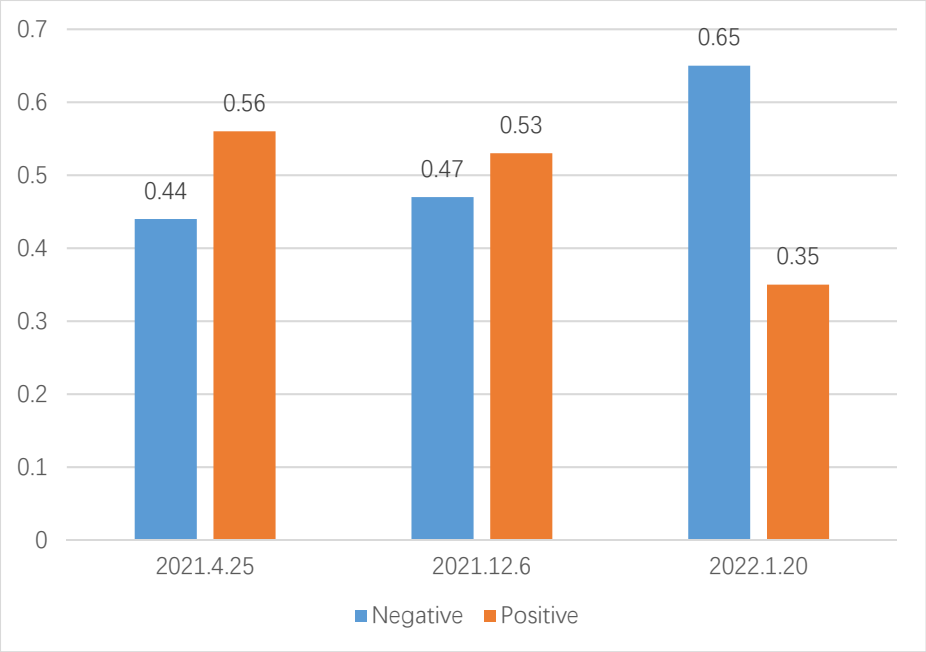

Figure 2. Emotional changes at different time points of the UK COVID-19

Globally, the COVID-19 epidemic began to break out at the beginning of 2020. Due to effective control and epidemic prevention measures in various countries, around April 2021, although COVID-19 had undergone several mutations, the global infection rate was still well controlled, and the epidemic situation in Britain was also well controlled. However, the emotion expressed in the video in April is negative, which may be caused by the spread of the global epidemic, the strong internal control in the UK and the poor life experience of People. In December 2021, the variant of COVID-19, Omikron, swept the world again. It was first detected in South Africa on November 9, 2021, and the overall risk of COVID-19 variant of Omikron was assessed as "very high" globally, which may be widely spread in the world. The emotional bias of users is also relatively similar, with negative proportion reaching 53%, lower than that in April, indicating that optimism is being released. In January 2022, even though Omikjon was widely spread, the people were confident after two years of anti-epidemic, so the user's mood was optimistic, and the positive mood had reached 65%.

3.2. Analysis of LDA Subject Extraction Results

We used LDA [6] to extract 10 subject words from each of the three topics of negative and positive comments in January 2022, as shown in Table 3.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202304_9(4).0011

Table 3. Negative and positive theme words of comment text

Negative			Positive		
topic0	topic1	topic2	topic0	topic1	topic2
news	canada	people	canada	canada	canada
people	people	england	time	follow	people
time	covid	boris	finally	people	god
boris	mask	covid	people	england	time
mask	time	canada	follow	god	news
australia	restrictions	mandates	sense	boris	countries
week	government	time	party	freedom	boris
lol	hope	mask	common	time	finally
global	masks	god	rest	country	christ
covid	follow	jesus	forget	government	england

Both negative and positive views, everyone pays attention to people, country (England, Canada), time, etc., but negative views also pay more attention to mask, covid, positive views pay more attention to party, Boris, god, etc., which shows that negative views pay more attention to the current epidemic situation and epidemic prevention measures, while positive views are more inclined to the policies of political parties, prime ministers, and religious beliefs.

In terms of details, among the negative views, topic0 focuses on the global scope, topic1 focuses on restrictions and government related policies, and topic2 focuses more on words with the nature of prayer such as god and jesus, which shows that everyone's views have three natures, the focus on epidemic policies in various regions, the focus on the epidemic prevention policies of their governments, and the prayer for the current epidemic. In the positive view, topic0 pays attention to common sense, hoping that everyone will follow the common sense of anti-epidemic and take a positive and optimistic view. topic1 pays attention to the government and Boris's policy of the lift of lockdown, topic2 pays attention to the religious words like god and charis, which shows that in the positive view, it also has three natures, the common sense of anti-epidemic, the lift lockdown policy of the national government and the Prime Minister, and the prayer for the current epidemic.

4. CONCLUSION

With regard to the epidemic situation abroad, many users have expressed their views or opinions on various platforms, including the release of emotions and worries about the epidemic, both optimistic and negative. In general, the public's concerns can be classified into four categories: epidemic situation in various regions, common sense of epidemic prevention, government epidemic prevention policy and lift lockdown policy, and prayer for epidemic situation. From the perspective of emotional analysis, the general mood of the British people is improving. Although the recurrence of the epidemic will still make the people feel negative, but after more than two years of anti-epidemic, the people's confidence in the government is positive, and they will also take the initiative to understand the common sense and policy of anti-epidemic as well as lift lockdown policy of the local government, and everything is going well.

REFERENCES

[1] He, J., Li, Y. (2022) Research on the emotional evolution characteristics of public opinion reversal events based on space-time perspective. Journal of Information Resources Management, 12(02):88-100.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202304_9(4).0011

[2] Xia, Y. (2019) The model and empirical study of "attenuation transfer" of netizens' emotion based on public opinion big data. Journal of Information, 38(3): 148-154.

[3] Rexiline Ragini J., Rubesh Anand P.M., Bhaskar V. (2018) Mining crisis information: a strategic approach for detection of people at risk through social media analysis. International Journal of Disaster Risk Reduction, 27: 556-566.

[4] Xiong, X., Li, Y., Qiao, S., et al. (2018) An emotional contagion model for heterogeneous social media with multiple behaviors. Physica A: Statistical Mechanics and Its Applications, 490: 185-202.

[5] Sundermeyer M., R. Schlüter, Ney H. (2012) LSTM neural networks for language modeling. Interspeech.

[6] Cao, R., Sun, M. (2020) Research on the improvement of public opinion comment text subject extraction based on LDA. Computer engineering & Software, 41(10):7.