International Core Journal of Engineering Volume 7 Issue 3, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0052

Processing of Unlabeled Data: Research on the emotion of Social Comment Text based on ANN

Junjie Li

School of Information Science and Technology, Chengdu University of Technology, Chengdu, Sichuan, 610051, China.

201813160612@stu.cdut.edu.cn

Abstract

The main research object of this paper is the emotion of unlabeled Weibo comments. The main research content is to effectively identify and label unlabeled Weibo comments. The main method used in this study is to use the python package jieba to extract feature values of comments, and then process and train the data based on ANN, so as to predict unlabeled comments and conduct tendency analysis.

Keywords

Unlabeled Weibo Comments; Python; Feature Extraction; ANN; Tendency Analysis.

1. Introduction

At present, the communication in the society are constantly developing. The number of Internet users has doubled every year, and the proportion of users in the world has increased from only about 0.6% in 1995 to 39% in 2014. The distribution of the world's population using the Internet has changed year by year. At present, the number of Internet users in China has accounted for about one-third of the total number of users in the world, and the proportion is still rising. The rapid development and rise of Weibo has made it an important application on the Internet after communication tools such as mailboxes and QQ. This kind of information interaction between people has become an indispensable way of communication in people's lives, and its influence is constantly expanding. It not only stays at the level of ordinary users, but also penetrates into enterprises, governments and other social organizations. In all aspects, due to the insufficient maturity and perfection of network supervision, the authenticity and credibility of the information on the network is not high, and there are many security risks such as the wanton spread of rumors and fake news. The network environment needs to be effectively cleaned up and improved, and the government are also paying close attention to the public's response to hot events. Online reports show that as of June 2015, there were 204 million people using Weibo in China, accounting for about 30% of the total number of users. Approximately 162 million communicated via mobile Weibo, which reached 27.3%, an increase of about 10% compared with 2014. The Internet is becoming more and more important to people's lives, and Weibo is becoming more and more important to people's daily lives. Therefore, how to judge the tendency of the views of the masses on the Weibo social platform has become one of the urgent problems in public opinion analysis.

At present, sentiment analysis is mainly divided into two ways: unsupervised and supervised. Most of the foreign research results focus on the emotional orientation of each user on Twitter. The literature [1-3] distinguishes the emotional orientation by matching certain rules, which mainly extracts the characteristics of the comment information, and divides the comment information according to different Situation division, recording the characteristics of comments, and dividing the data set into categories according to rule matching, without using machine learning and other methods.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0052

Liu B.[4] builds a dictionary library through knowledge of related fields, and wants to realize the comment analysis between different fields in a specific way, and further develops in cross-field research, and obtains a better classification effect. Turney [5] judges the sentimental tendency of the text by the contribution value in the dictionary, but the effects and rules are difficult to be compatible, so it is not advisable to rely too much on the sentiment dictionary. As for the unsupervised learning of sentiment dictionaries, the model designed by Neviarouskaya et al. [6] has been used early in foreign countries. It has achieved better results for simple sentence comments on Twitter, but for complex sentence patterns, the sentiment dictionary constructed by manual selection Manpower is limited and the effect is very unsatisfactory. It can be seen that the use of the emotional dictionary alone is too one-sided. Prasad et al. [7] used Bayesian machine learning methods to classify blogs and achieved certain results. This is the earlier application of machine learning to emotion classification. Pang et al. [8] judged the emotional value of related movie reviews in English, and its effect was more significant in the early stage. Davidov et al. [9] manually annotated the content of the text in Twitter, extracted the topic tags and emoticons, and realized emotion classification through the K nearest neighbor algorithm classifier. Go et al. [10] realized the use of distance learning methods to form training data, and obtained better results through topic and keyword analysis combined with corresponding learning methods. Under the conditions of continuous maturity in learning methods, Li S. [11] divided sentences into subjective and objective categories, and tested them through supervised learning. At the same time, they used semi-supervised learning methods and compared them to obtain the Good effect of sentence subjective and objective classification. Jiang L et al. [12] treated the problem of Tweets text classification as a general classification problem, used machine learning to classify subjective and objective content, and achieved its classification goals by extracting some corresponding features. Barbosa et al. [13] compared various features of Tweets and extracted features according to their weights, and obtained the ranking of feature weights, which made a great contribution to feature selection. Agarwal et al. [14] chose to focus on the characteristics of Twitter itself. Since spoken language and symbols occupy a certain proportion, they extracted their own characteristics in a targeted manner. WilsonT. [15] proposed his own method for sentiment analysis in the form of phrases, and judged sentiment orientation through his own classifier. Regarding the particularity of Twitter, Mohammad et al. [16] carried out a series of engineering constructions on characteristics, which were highly targeted and integrated various advantages, and had good engineering for the analysis of comments on Twitter. judgment.

In the work, This topic proposes an analysis of the tendency of unlabeled Weibo comments based on ANN, which aims to quickly and accurately judge the tendency of comment text. Therefore, the government and other departments can monitor public opinion based on the flow of judgment results, provide certain decision-making methods, conduct certain psychological counseling to the crowd, prevent excessive behavior and prevent violent incidents

2. Weibo comment sentiment analysis

2.1 Weibo data collection

Use python to crawl data on Weibo, by inputting keywords, such as violent vocabulary: idiot; active vocabulary: happy and other characteristic words to crawl. The final collected data is 7224 characteristic word comments.

2.2 Data cleaning and manual labeling

There are a lot of rubbish and useless blog posts in Weibo comments. Although this type of blog post has some characteristic words, it cannot be used for emotional analysis. Such blog posts will interfere with the results of emotional analysis and should be deleted.After data cleaning, there are 4,715 useful comments, of which 2,376 are negative comments and 2,339 are positive comments. Part of the data is shown in Table 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0052

Table 1 Weibo comments and sentiment analysis

comment	Emotional orientation	Data annotation
To be honest, if you encounter such a stupid, you will have to fight for your life.	negative	1
Accounting, I love you!!!!	positive	-1
Happy first day of class	positive	-1

2.3 Building a label prediction model based on ANN

2.3.1 Sample data feature value extraction

The text format of Weibo comments is not fixed and the content is extremely diverse, making it difficult to directly apply it to model training. Therefore, the original Weibo comment data needs to be preprocessed, that is, feature value extraction. This article mainly uses jieba to extract the following feature values from the original data:

feature value 1: The number of derogatory words;

feature value 2: The number of words containing blood and violence;

feature value 3: the number of unfriendly fixed sentence structures.

feature value 4: The number of punctuation marks containing exclamation marks and other strengthening tone;

By extracting the number of feature values and the emotional tendency of their annotations as data for ANN training, the text format is converted into pure data content, which greatly reduces the difficulty of training.

2.3.2 Predictive model building and training

The ANN (artificial neural network) structure is used to build the label prediction model of this topic. Figure 1 shows the structure of the model. The model consists of an input layer, a hidden layer and an output layer. Each neuron in one layer is connected to all neurons in the adjacent layer, so the entire structure is similar to a network. In terms of function, the input layer does not require any data processing, and only inputs data from the outside to the network. Complex data processing is carried out by the hidden layer and the output layer. The specific processing process is shown in Figure 2. The training process of the neural network model is to iteratively adjust the connection weights between neurons and the deviation of neurons until the specific requirements are met. The input vector of the model is the eigenvalue vector composed of the above eigenvalues [Sv1,Sv2,Sv3,Sv4,Sv5], the output vector is the value of the degree of tendency, the value range is [-1,1], 1 means the most negative tendency, -1 indicates the most positive tendency, 0 means neutral speech.

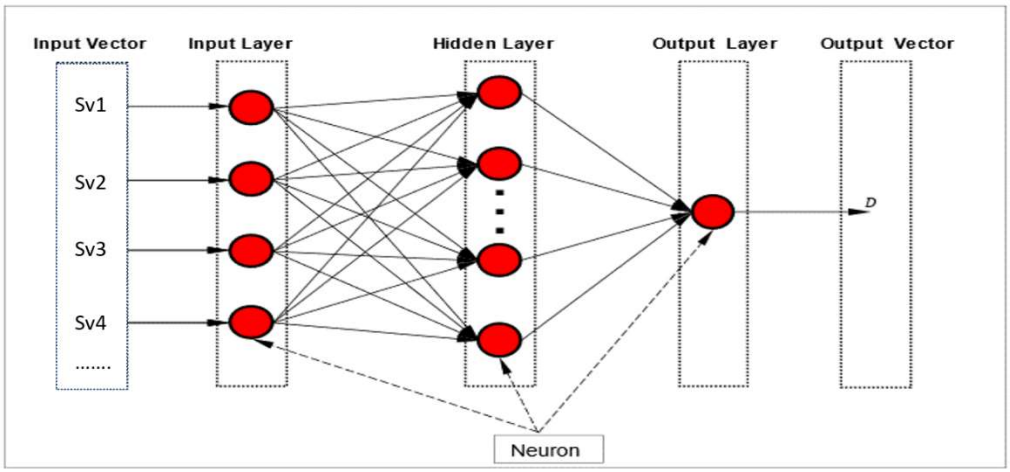

Figure 1: Based on the ANN prediction model structure

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0052

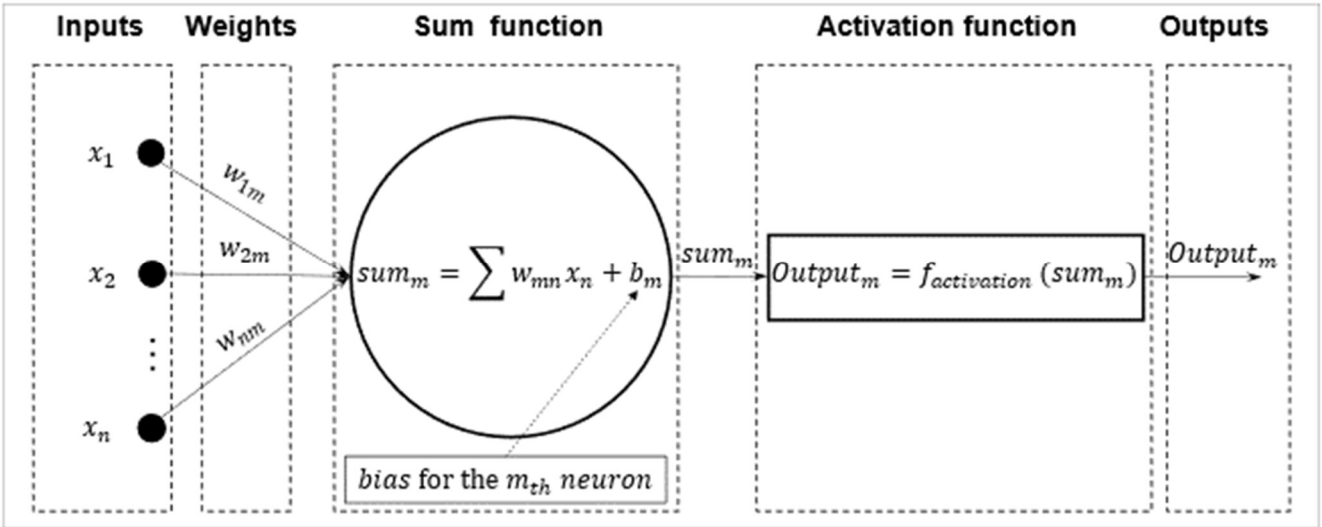

Figure 2: Data processing process of hidden layer and output layer unit

According to the above model structure, the training process is as follows:

Step 1: Set the parameters of the network model. Table 2 lists some important parameters and their values.

Table 2: ANN model configuration

parameter	value
Number of hidden layers	1
Number of neurons in the input layer	4
Number of hidden layer neurons	9
Number of neurons in the output layer	1
Hidden layer activation function	Tan-Sigmoid (tansig)
Output layer activation function	Linear (purelin)
Target mean square error	1.5e-4
The maximum number of iterations	1000
Learning rate	adaptative
Training algorithm	Levenberg-Marquardt Backpropagation []

Step 2: Assign initial weights to all connections, and assign biases to all neurons in the hidden layer and output layer.

Step 3: Input the input vector [Sv1, Sv2, Sv3, Sv4, Sv5] into the ANN model.

Step 4: According to the input vector in step 3, calculate the output values of all neurons in the hidden layer and the output layer by formula 1. The output of any neuron in the input layer is equal to the input of that neuron.

$$Output_m^k = f_{activation}(sum) = f_{activation}\left(\left(\sum_{n=1}^{n=N} w_{mn} \times Output_n^{k-1}\right) + bias_m^k\right) \quad (1)$$

($Output_m^k$ represents the output of the mth neuron in the current layer $Output_n^{k-1}$ represents the output of the nth neuron in the previous layer; N represents the number of neurons in the previous layer, W_{mn} represents the mth neuron in the current layer and the nth neuron in the previous layer Element connection weight value; $bias_m^k$ is the deviation of the mth neuron in the current layer; $f_{activation}$ represents the activation function)

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0052

Step 5: Obtain the output value of the only neuron in the output layer according to Step 4, and calculate the error and square error between the output value ($D_{calculated\_i}$) and the target output value (D_{actual_i}) by formulas 2 and 3 respectively.

$$E_i = D_{calculated\_i} - D_{actual_i} \tag{2}$$

$$SE_i = \left(D_{calculated_i} - D_{actual_i}\right)^2 \tag{3}$$

(E_i and SE_i respectively represent the error and square error between $D_{calculated\_i}$ and D_{actual_i}; $D_{calculated\_i}$ and D_{actual_i} are the calculated output value and target output value of the only neuron in the output layer corresponding to the i-th input vector.)

Step 6: Adjust the weights and biases according to the Levenberg-Marquardt backpropagation algorithm. The algorithm provides weight adjustment rules as shown in formulas 4 and 5.

$$wb_{new} = wb_{old} + [J^T J + \mu I]^{-1} J^T E_i \tag{4}$$

$$J = \left[\frac{\partial E_i}{\partial wb_i} \cdots \frac{\partial E_i}{\partial wb_P}\right] \tag{5}$$

(wb_{new} represents the current weight deviation matrix; wb_{old} represents the weight deviation matrix before updating; J represents the Jacobian matrix; P represents the number of elements in the weight deviation matrix; I represents the identity matrix; μ is the adjustable factor, which is affected by SE_i)

Step 7: Repeat the steps from step 3 to step 6, and observe whether the current mean square error (MSE) of all input vectors calculated by formula 6 is less than the allowable error. If yes, go to step 8. Otherwise, repeat from 3 to 7 until the number of repetitions reaches the maximum number of iterations.

$$MSE = \frac{\sum_{i=1}^{K}\left(D_{calculated\_i} - D_{actual_i}\right)^2}{K} \tag{6}$$

(K is the number of input vectors in the training data set or the size of the training data set)

Step 8: Stop the iteration, and all model parameters (including all weights and deviation values) have been fixed.

Step 9: Validate the model using the validation data set, and calculate $MMSE_{validation}$ through the following formula 7. If $MMSE_{validation}$ does not exceed the allowable error, it proves that the neural network is established successfully. Otherwise, it means that the model is overfitting, and then the model parameters should be changed and all the steps above should be repeated.

$$MMSE_{validation} = \frac{\sum_{i=1}^{i=\mu}\left(R_{predicated\_i} - R_{a\_target\_i}\right)^2}{N} \tag{7}$$

($\mathrm{MMSE}_{\mathrm{validation}}$ is the average MSE between all outputs of all input vectors and all target outputs in the verification data set; N is the size of the verification data set.)

Label unlabeled data based on the model

According to the label prediction model built in step 2, the label can be predicted for unlabeled Weibo comments, which becomes labeled data and added to the labeled data set.

Build the final prediction model based on all label data

Table 3: Model result prediction

Number of training sets	Number of prediction sets	Number of test sets
3300	707	707

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0052

According to all available labeled data sets, the final Weibo comment tendency prediction model is constructed. The implementation method is basically the same as the research content (2), but the training algorithm in step 6 will be optimized to achieve a faster convergence rate.

From the final model test incorrect rate, we can see that the incorrect rate is 3.66%. It is concluded that the prediction result of this model is accurate.

3. Conclusion

In the past, prediction models were built by training on existing labeled data sets, but the capacity of labeled data sets is generally limited, and the accuracy and reliability of prediction models based on limited data sets are not high. This topic first transforms unlabeled data into labeled data, and then forms a huge labeled data set to participate in the training of the model, which greatly improves the accuracy and reliability of the final model.

References

[1] Joshi A, Balamurali A R, Bhattacharyya P, et al. C-Feel-It: a sentiment analyzer for micro-blogs, 2011[C]. Association for Computational Linguistics, 2011.

[2] Das A, Bandyopadhyay S. Dr Sentiment knows everything!, 2011[C]. Association for Computational Linguistics, 2011.

[3] Chesley P, Vincent B, Xu L, et al. Using verbs and adjectives to automatically classify blog sentiment[J]. Training, 2006,580(263):233.

[4] Liu B. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1):1-167.

[5] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, 2002[C]. Association for Computational Linguistics, 2002.

[6] Neviarouskaya A, Prendinge H, Ishizuka M. SentiFul: A lexicon for sentiment analysis[J]. Affective Computing, IEEE Transactions on, 2011,2(1):22-36.

[7] Prasad S. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods[Z]. Technical Report, 2010.

[8] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2):1-135.

[9] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags an smileys, 2010[C]. Association for Computational Linguistics, 2010.

[10] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, Stanford, 2009,1:12.

[11] Li S, Huang C, Zhou G, et al. Employing personal/impersonal views in supervised and semi-supervised sentiment classification, 2010[C]. Association for Computational Linguistics, 2010.

[12] Jiang L, Yu M, Zhou M, et al. Target-dependent twitter sentiment classification, 2011[C]. Association for Computational Linguistics, 2011.

[13] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data, 2010[C]. Association for Computational Linguistics, 2010.

[14] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data, 2011[C]. Association for Computational Linguistics, 2011.

[15] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis[J]. Computational linguistics, 2009,35(3):399-433.

[16] Mohammad S M, Kiritchenko S, Zhu X. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets[J]. arXiv preprint arXiv:1308.6242, 2013.