Machine Learning Based Sentiment Analysis of Message on Twitter

Mingyou Dai*

Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Macau, 999078, China

* Corresponding Author Email: 1210017784@student.must.edu.mo

Abstract. In the age of information explosion, it is a very important and challenging task to extract the required features from a huge amount of information. The emotion expressed by the text information is one of the most important features of the information. However, there is not much research in this field, so it is of great significance and exploratory to the text sentence emotion analysis. In order to compare and explore better feasibility, both the sequential neural network model and the random forest model were built. Through the contrast between the two models, the machine learning of datasets composed of Twitter comments is carried out to analyze the emotion of the text, and emotion extraction is the research topic of this paper. In this paper, emotion analysis was studied in the order of data processing, model building, and result analysis, and an accuracy of about 90% was finally achieved, which is a good result, from which it can be seen that the constructed neural network model plays a good role.

Keywords: Machine learning; semantic analysis; neural network.

1. Introduction

In the era of the Internet's explosive growth, everyone has their own opinions in the face of all kinds of information, and people often express their views with emotions to form comments, which constitutes a huge amount of information on the Internet. In this huge amount of information, each comment is a medium for emotional expression. They reflect people's thoughts and expressions, so it is very necessary to extract and analyze the emotions contained in it.

In recent years, people have started to pay greater attention to emotion analysis. Study of text information's emotional content is one of the most important research fields in emotion learning. Many applications are also mining the value of text emotion analysis [1, 2], which can help the application more accurately understand the user's preferences and habits, help it better operation, and create a better user experience.

In addition, sentiment analysis is creating commercial value for many companies. In the field of emotion computing and emotion analysis, the public's interest in their products can be predicted by collecting users' emotion information [3, 4], which is of great significance for companies to formulate marketing strategies and future product research and development orientation. The commercial application of tools to mine real-time network emotions is also attracting more and more companies to invest.

The realization of emotion analysis is also of great use to the society. For example, government departments can analyze people's online comments based on emotion to judge people's emotional orientation [5, 6]. They can timely understand people's reactions, sort out their attitudes and understand their voices, so that people's thoughts can be paid attention to in time. In addition, it can also find out the bad direction of network public opinion and bad emotions in time, give them correct guidance, and maintain the long-term peace and stability of the society.

From the above applications, it can be seen that the emotional analysis of text messages is worth studying, but extracting them is not an easy task. Each person's language has its own way of expressing itself, and it is impossible to process such a large amount of information manually because it would be too laborious. In order to solve this problem. To solve this problem, an attempt was made to find a faster and cheaper way to make sentiment analysis possible. Nowadays, the development of

computers helps people to solve problems that are not easy to solve by humans. The analysis of emotion can be inferred based on the sentence's word and phrase meanings, so the extraction of emotion from the words in the sentence is very important in analyzing the overall emotional orientation [7]. Machine learning has the advantage of feature extraction, obtaining experience from existing data to optimize learning efficiently and autonomously [8]. In view of the various characteristics of comment text, machine learning is a good choice for sentiment analysis.

Therefore, machine learning based on Twitter comments for mood analysis is chosen to achieve the purpose of learning emotion judgment analysis. In Twitter comments, each person has a different view of the tweet, which also contains different emotions [9]. A datasets of Twitter comments is selected, where 0 and 1 are used as labels to correspond to bad and good sentiments, which was chosen to conduct the study and test of machine learning. Through emotion analysis, judgment of the overall proportion of positive and negative emotions can be applied to make the large-scale emotional proportion of comments visualized and intuitive. This paper uses two models to compare machine learning, focusing on the analysis of the model with better accuracy.

The following chapters are arranged as the method part to introduce the data collection source and processing, as well as the specific method implementation of the two models. The result part uses two different verification indicators.

2. Method

This chapter introduces how to collect the datasets and how to process data firstly. Then the methods for processing data and machine learning are introduced, and finally the precision rates are introduced. It can be visualized in Figure 1.

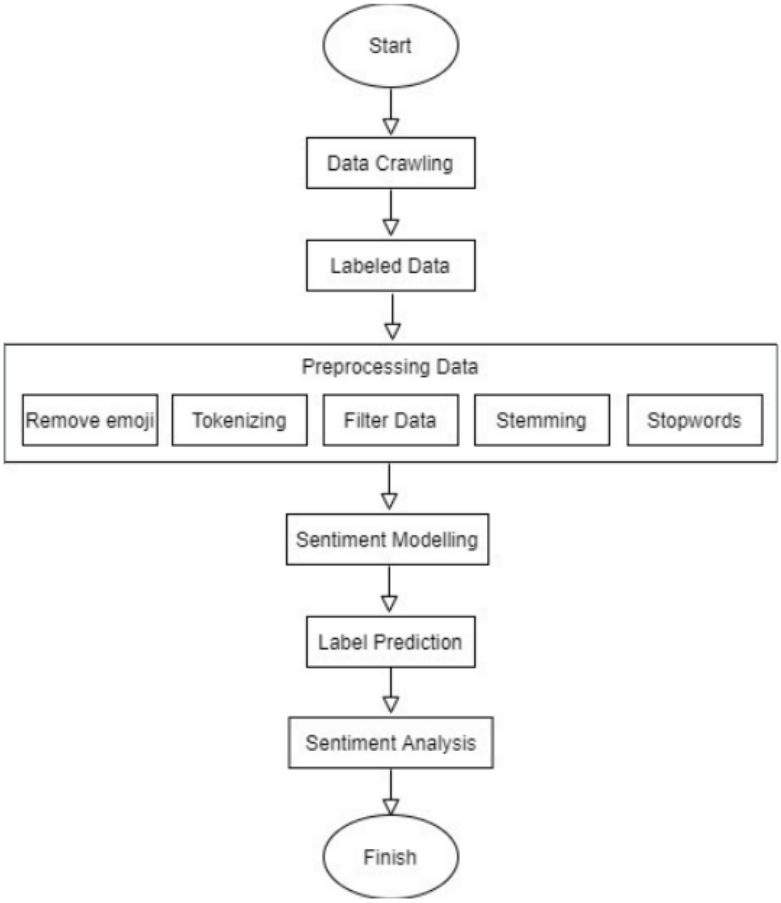

Fig. 1 The emotional analysis flow chart

2.1. Data

2.1.1 Data Introduction

The data [10] comes from Twitter. 50,000 comments from Twitter users with emotional characteristics are extracted. 40000 comments were included for the mainly training set, and used 5000 comments for the second part of the training, as well as the 5000 comments for the test sets. There are two sections to the data: text and label. Text is the content of the comment, and label is the emotion label corresponding to the comment, where 0 represents negative and 1 represents positive. In these three sets, the distribution of label 0 and 1 (negative comments and positive comments) of data is basically average.

2.1.2 Data Processing

By converting uppercase letters to lowercase letters, deleting punctuation, numbers, rare characters, and other simplified statements, it is convenient for later training and vectorizing the data. Data vectorization, using the count vectorizer in sklearn, is carried out by using the frequency of each word in the simplified statement, so as to better extract the features of the text and digitize them. Table 1 below is an example of data before data processing, and Table 2 is an example of data after data processing.

Table 1. Data before data processing

Text	Label
And then comes the most dishonest cheat ending I've seen, much worse than"April Fool's Day" - where at least it made sense.	0
The second film about the adventures of the Gaulois pair Asterix & Obelix is 10 times better than the first.	1
The sound is terrible, going from too low to hear the conversations, to blaring sound in seconds. The plot is absolutely implausible, the acting is mediocre.	0
The acting is superb, specially from the stepmother and the main girl. Those two are worth the price of the ticket alone. Do yourself a favor and watch this awesome film.	1

Table 2. Data after data processing

Text	Label
comes dishonest cheat ending ive seen much worse than april fool day least made sense	0
second film adventure gaulois pair asterix obelix time better first	1
sound terrible going low hear conversation blaring sound second plot absolutely implausible acting mediocre	0
acting superb specially stepmother main girl two worth price ticket alone favor watch awesome film	1

2.2. Method Implementation

Then there's the training. The sequential neural network model in TensorFlow is mainly used, and then random forest are used for comparison.

2.2.1 The Sequential Neural Network Model

So, what is a neural network? The human neural network served as a model for artificial neural networks, which has a significant number of neurons, and the interconnection of neurons forms a highly complex and flexible dynamic network. Therefore, people have also studied the leverage of an artificial neural network and computer realization to create a computing model, which is constructed with a number of nodes directly related to each other. Each node is calculated through a specific operation, also known as the excitation function, and output as the input of the next node.

In the sequential neural network model in TensorFlow, two fully connected layers have been used. The fully connected layer can combine the review features into a single output value, which greatly

reduces the influence of feature location on classification and ignore spatial structures that are not necessary for statement analysis to better classification.

Firstly, two fully connected layers constructed ReLU and sigmoid as activation functions respectively. Next, the loss function is set as binary cross entropy, the optimizer is Adam, and the network evaluation index is marked as the training method of accuracy. Finally, for training, Input the text and label vectorized by the training set consisting of 40,000 data respectively, and then use the second set of 5000 data to form the datasets as the test training. The epoch number is set to 10, and the batch size is set to 32. By comparing the loss and accuracy of different epochs, the best epoch can be found, and calculate its accuracy through a test set composed of 5000 data. After that, the learning effect was shown by drawing confusion matrix, ROC curve and manually entering statements for testing.

2.2.2 The Random Forest Model

For random forest, numerous trees make up the random forest, all of which are unrelated to one another. The integrated algorithm based on the decision tree classifier can obtain the final prediction result through voting of these independent decision trees. Compared with a single decision tree, it has the ability to randomly extract samples and features, which makes it have a higher ability to resist overfitting. And its integrated algorithm also makes it more accurate. The same data preprocessing (vectorized) like sequential neural network model is used for random forest model, The model is constructed by random forest classifier, and the maximum accuracy is found by changing the estimator value from small to large. Finally, the confusion matrix and ROC curve were drawn to display the results, and the manual input sentences were tested.

2.3. Evaluation Index

For the purpose of assessing the model's learning capabilities, the confusion matrix and ROC curve are introduced. The confusion matrix can clearly show the gap between the prediction and the actual in a graph, which can intuitively evaluate the prediction effect. The true positive rate is represented by the ordinate of the ROC curve, while the false positive rate is represented by its abscissa. These two evaluation indexes can be more intuitive to see whether positive and negative judgments of sentence meaning are good through training.

3. Result

To ensure that the experiment was successful, both models pass the manual input sentence test and the confusion matrix and ROC curve are used to show the experimental results.

The confusion matrix for the sequential neural network model is shown in the bottom plot in Figure 2, while the confusion matrix for the random forest model is shown in the second plot in Figure 3.

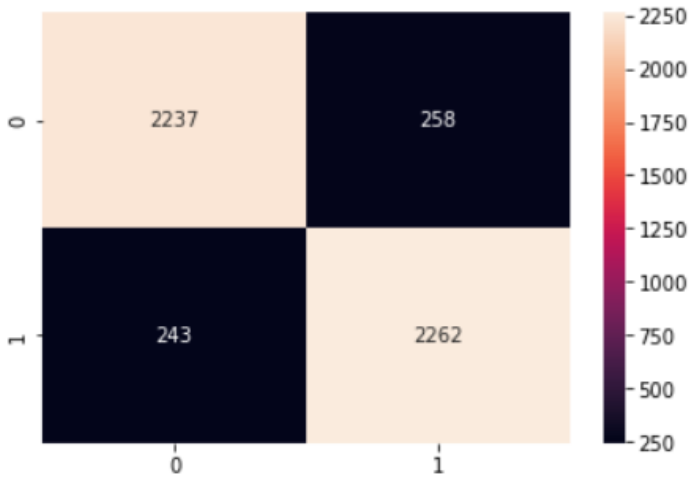

Fig. 2 the confusion matrix of the sequential neural network model

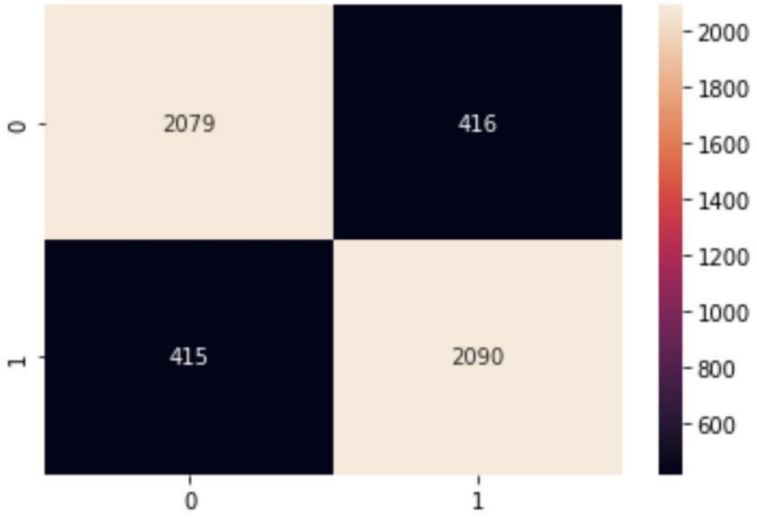

Fig. 3 the confusion matrix of the random forest model

In the labels, a negative emotion is represented by 0, and a happy emotion is represented by 1, where the first row is the negative of the predicted label and the second row is the positive of the predicted label. The first column is the negative of the real label and the second column is the positive of the real label. As a result, row 1, column 1 represents a true negative, row 1, column 2 a false negative, row 2, column 1 represents a false positive, and row 2, column 2 represents a true positive. Accuracy, recall, and precision can be calculated from the confusion matrix. The total of true positive and true negative labels divided by the total number of labels equals accuracy. Recall is calculated as true positive divided by the total of true positive and false negative. The ratio of true positives to the total of both true and false positives is known as precision. On the basis of this, it can be determined that the sequential neural network model's accuracy is 89.98%, recall is 89.76%, and precision is 90.30%. Additionally, the random forest model's accuracy is 83.38%, recall is 83.40%, and precision is 83.43%.

The ROC curve for the sequential neural network model is shown in Figure 4 of the plot below, and the ROC curve for the random forest model is shown in Figure 5.

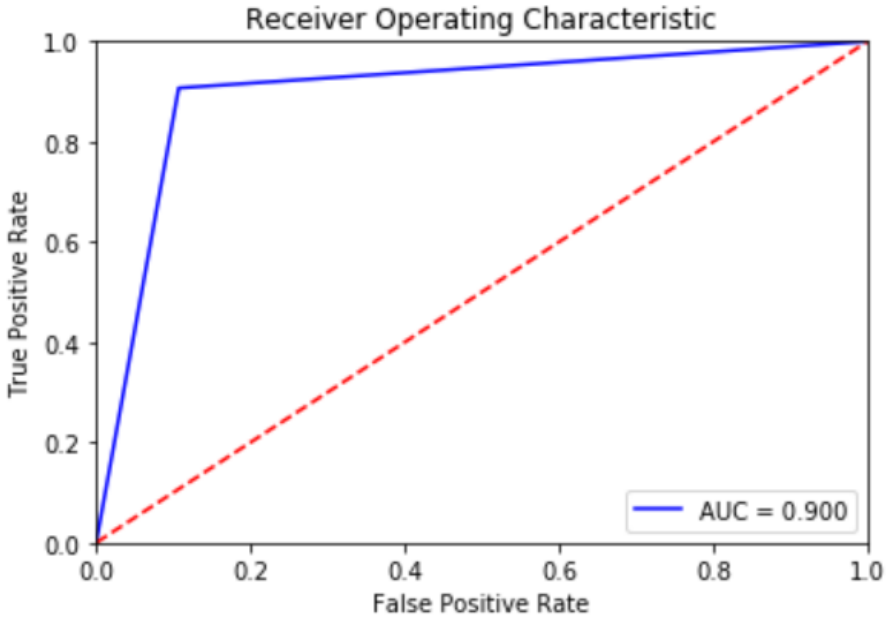

Fig. 4 the ROC curve of the sequential neural network model

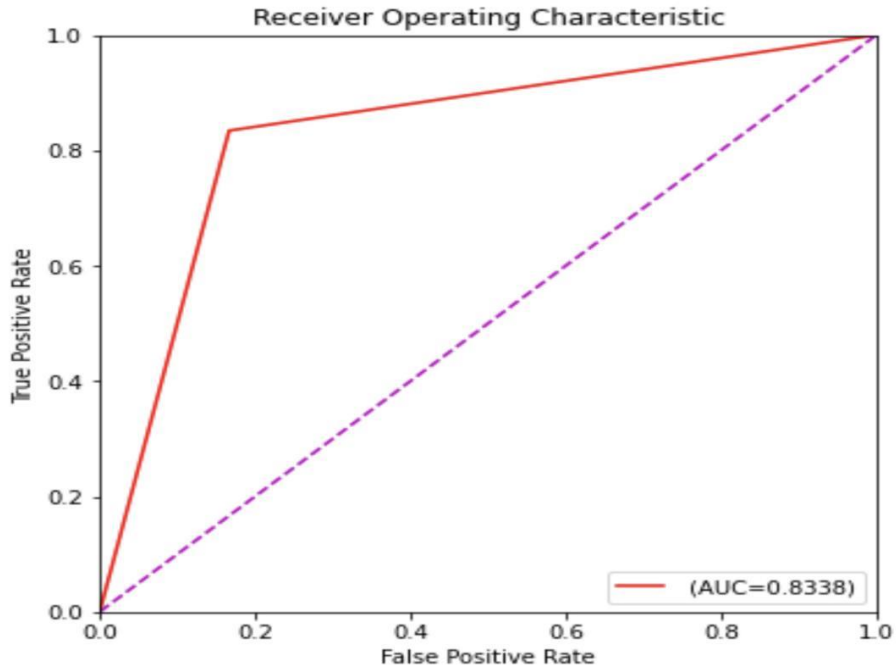

Fig. 5 the ROC curve of the random forest model

The sequential neural network model's and random forest model's curves are higher than the diagonal line, which indicates that the learning effect of the two models is reasonably strong, according to the ROC curve. It is clear that while the accuracy of the random forest model is 83.38%, that of the sequential neural network model is roughly 90%.

In summary, the learning performance of the sequential neural network model is better than that of the random forest model in this study.

4. Discussion

This paper addresses the problem of emotion classification using machine learning from Twitter comments. The overall accuracy of the two models is good, and the sequential neural network model can achieve an accuracy of about 90% More complex neural networks are recommended because it can be seen that in the sequential neural network model, the highest accuracy is reached when epoch reaches 2. It can see that this work may get better results by using more complex neural networks to achieve more epochs, which is a point that can be improved. In addition, training and testing with more data may also help improve the performance.

5. Conclusion

Based on Twitter datasets, this paper conducted data processing by simplifying and deleting words in sentences, and extracted features for model comparison analysis. In this paper, confusion matrix and ROC curve are used to compare models and show model performance. In sequential neural network model, the accuracy is 89.98%, recall is 89.76%, and precision is 90.30%. In random forest model, accuracy is 83.38%, the recall is 83.40%, and the precision is 83.43%. When the sequential neural network model and the random forest model were compared, the sequential neural network model outperformed the random forest model. It can be seen that in the statement emotion classification studied in this paper, the deep learning model using neural network is superior to the random forest model, the traditional machine learning method. In other words, in the end, the best results came from the sequential neural network model, the average accuracy is very close to 90 percent. The realization of sentence emotion analysis can provide help for the addition of social software comments emotion proportion module. It also can analyze users' emotions in business operation to get the direction of advertising, product research and development, etc., which can help

it gain more profits. Furthermore, it can help supervise the guidance of online comments and maintain network security. Of course, the accuracy of testing with different data sets is bound to vary, so using more data for training and testing will certainly be of great help to improving the model. For example, the data set is in English, so in the future, the data can be expanded to more languages, such as Chinese, Arabic and so on, which will definitely play a bigger role. However, due to the gap between different language expressions, the data processing method may need to be changed, which is also a big difficulty for future research and testing. I believe there will be better models and data processing methods to deal with these difficulties in the future.

References

[1] Wang X, Kou L, Sugumaran V, et al. Emotion correlation mining through deep learning models on natural language text. IEEE transactions on cybernetics, 2020, 51(9): 4400-4413.

[2] Gautam G, Yadav D. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh international conference on contemporary computing (IC3). IEEE, 2014: 437-442.

[3] Cambria E, Das D, Bandyopadhyay S, et al. Affective computing and sentiment analysis. A practical guide to sentiment analysis. Springer, Cham, 2017: 1-10.

[4] Cambria E, Poria S, Hussain A, et al. Computational intelligence for affective computing and sentiment analysis. IEEE Computational Intelligence Magazine, 2019, 14(2): 16-17.

[5] Mejova Y. Sentiment analysis: An overview. University of Iowa, Computer Science Department, 2009.

[6] Habimana O, Li Y, Li R, et al. Sentiment analysis using deep learning approaches: an overview. Science China Information Sciences, 2020, 63(1): 1-36.

[7] Karamibekr M, Ghorbani A A. Lexical-syntactical patterns for subjectivity analysis of social issues. International Conference on Active Media Technology. Springer, Cham, 2013: 241-250.

[8] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349(6245): 255-260.

[9] Bollen J, Pepe A, Mao H. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. arXiv preprint arXiv:0911.1583, 2009.

[10] Maas A, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011: 142-150.