Social Sentiment Analysis Using Classifiers and Ensemble Learning

To cite this article: Linxiang Zhang 2019 J. Phys.: Conf. Ser. 1237 022193

View the article online for updates and enhancements.

You may also like
- Sentiment Analysis Of Movie Reviews Based On Improved Word2vec And Ensemble Learning
  Xiaohan bao, Shasha Lin, Ruilin Zhang et al.
- Transmission characteristics of investor sentiment for energy stocks from the perspective of a complex network
  Yajie Qi, Huajiao Li, Nairong Liu et al.
- Image recognition using convolutional neural network combined with ensemble learning algorithm
  Weilong Mo, Xiaoshu Luo, Yexiu Zhong et al.
Social Sentiment Analysis Using Classifiers and Ensemble Learning

Linxiang Zhang
Internation School, Beijing University of Posts and Telecommunications, Beijing 100876, China
zhanglinxiang@bupt.edu.cn

Abstract: Instant message applications has become the trend of present social network for connecting people nowadays. With such massive information transmitted between end users, it may provide valuable hidden knowledge for data analysing and help us to obtain some insights of how the messages can influence current social trends. During the analyses, the author uses several different classifiers to conduct sentiment collection and use Ensemble Learning to aggregate the results. As a result, the author maps data about users’ sentiment graph in database of given category.

1. Introduction
Classification is significant in the real world and is widely used in text classification, image classification. In recent years, many classification methods have been developed, such as the Naïve Bayes, Logistic Regression, Decision Tree and Support Vector Machine. WhatsApp generates a huge amount of data. First, the author use LDA model to analyzes the content and divided people into different groups. Then, the authors analyze sentiment of people in groups. People can express their own emotions when talking with each other. The author tries to seek out whether users express positive or negative feelings. By using five kinds classifiers to do the analysis and using voting classifier to aggregate the results from these five classifiers the author gets the final results. These five classifiers are Naive Bayes classifier, Linear SVC classifier, Logistic Regression classifier, MNB classifier and Bernoulli NB classifier. These classifiers use different methods to do the analysis. Then the author uses Ensemble method which aims to compose a meta-classifier with different classifiers. Compared with single classifier, the meta-classifier is more efficient and has better generalization performance. By using the voting classifier, the results predicted by most classifiers are used as the final class label. Then the author analyzes the sentiment for people in different groups and get their emotions changes. A few people in some groups may be extremely active. However, a number of people in some groups may be not such active. Therefore, the author draws a popularity graph to show the five most popular groups in different time.

2. Background
2.1 Five classifiers used in sentiment analysis
In this paper, the author analyzes the sentiment of people in these public groups mainly using Ensemble method. The author uses five classifiers that are Naive Bayes classifier, Linear SVC classifier, Logistic Regression classifier, MNB classifier and Bernoulli NB classifier. Voting classifier is a kind of classifier to aggregate the results from classifiers and get the final results. In these classifiers, Gaussian NB classifier, Multinomial NB classifier, Bernoulli NB classifier are all belong to Naive Bayes Classifier.
[1] The author will introduce these classifiers later.

2.1.1 Naive Bayes
Bayesian rule developed by the English mathematician Thomas Bayes to describe the relationship between two conditional probabilities

In machine learning, the two most widely used classification model are Decision Tree Model and Native Bayesian Model. Naive Bayes Classifier has a solid mathematical foundation and stable classification efficiency. The Naive Bayes Classifier is relatively simple since it is based on a simple assumption that attributes are conditionally independent of each other when the target value is given. For the given items to be classified, the probability of each category appearing under the condition of this item appearing and which one is the largest is considered as which category the item to be classified belongs to. There are three kinds of classifiers in Naive Bayes Classifiers which are Gaussian NB classifier, Multinomial NB classifier, Bernoulli NB classifier. [2]

2.1.2 Gaussian NB classifier
When dealing with continuous data, a typical assumption is that the continuous values associated with each classify are distributed according to Gauss distribution. Therefore, the author can convert some continuous variables into discrete values. Assuming that the train set contains continuous value \( x \), we classify the data according to categories and calculate the mean and deviation of each classification.

2.1.3 Multinomial NB classifier
For a multinomial distribution event model, the sample represents the frequency of a specific event, which is generated by multinomial distribution. This is a typical event model for document classification, where events represent the occurrence of a word in a single document, and the value represents the number of some word appearance. If the given class and feature do not appear together in the training set, the probability estimation based on frequency is 0. When they multiply with each other, all information in other probability will be eliminated. Therefore, researchers usually set a small sample correction called pseudo count in all probability estimates to prevent estimated probability to be set to 0. This kind of method is called Laplace smoothing. When the pseudo count is set to 1, the Lidstone Smoothing is used in usual.

2.1.4 Bernoulli NB classifier
In the multivariate Bernoulli event model, the feature is two variables which describe the input. It uses global features and each eigenvalue is Boolean in usual. This model is usually used for text categorization, which uses two features instead of word frequency. This event model is particularly suitable for short text classification.

2.1.5 Logistic Regression Classifier[3]
Regression is actually the estimation of unknown parameters of known formulas. In fact, Logistic Regression is a linear regression normalized by logistic equation. In Logistic Regression Classifier, a part of a given training sample is marked as 1 (positive), and the rest is marked as 0 (negative). Train a classifier from these samples and predict the probability of future outcomes based on the performance of historical data.

2.1.6 Linear SVC classifier
Support Vector Classifier searches the best possible linear classifier according to the distribution of training samples. Samples that determine the location of the classification boundary are not all training data, but two data points of different categories with the smallest spacing between the two categories, namely "support vector". Thus, a few training samples which are most effective for prediction tasks can be screened from massive or even high-dimensional data. This kind of classifier is similar to SVC with parameter kernel="linear", but it is implemented lib linear not lib SVM. Therefore, it has more flexibility
in the choice of penalties and should scale better to large numbers of samples.

2.1.7 Voting Classifier [4-5]
The idea of Ensemble Learning is: one problem (such as classification problem) involves many kinds of algorithms in prediction (the algorithm in the following figure can solve classification problem). Among the prediction results, the prediction category that appears most is chosen as the final prediction category of the sample. In voting classifier, the results of various models are aggregated by voting. The results of the selected models can be good or bad, so the result of aggregation is the best. The basic idea of voting Classifier is to combine different machine learning models and predict the final labels by voting or averaging. For example, there are five classifiers, classifier1 → class2, classifier2 → class2, classifier3 → class3, classifier4 → class4, classifier5 → class5, then the final predict class is class 2.

3. Design and Implementation

3.1 Sentiment Analysis

3.1.1 Method for the analysis
The author uses five classifiers to do the analysis of sentiment and use voting classifier to aggregate the results from the five classifiers and get the final result. The five classifiers are Gaussian NB classifier, Multinomial NB classifier, Bernoulli NB classifier, Logistic Regression classifier, Linear SVC classifier. These five classifiers use different methods to do the classification. Voting classifier select the result with the most votes as the final results. For each classifier, the author gets the corresponding learning models and store them in the files so that following analysis can use these files.

3.1.2 Training dataset and test dataset
These five classifiers are supervised learning. Therefore, the author should input the training dataset and test dataset which is label data to build the learning model for each classifier and use the models to analyze the new data from WhatsApp public groups. The author gets these two kinds of datasets containing positive data and negative data from Standford sentiment analysis data. Figure 2 shows the datasets the author uses for the training and testing.

3.2 Popularity Analysis
A number of people in some WhatsApp public groups are extremely active and send much more message in the groups. However, a few people in some other groups are not such active and the volume of groups’ message is much less. Therefore, the author tries to find the popularity of these groups and finds the five most popular groups in different time.

The author makes statistics for the volume of message for each group. When one user sends a piece of message, the author will increase one for the volume of message of corresponding group. Then the author gets all Groups’ volume of message and obtains the five most popular groups. The author draws the final results in a bar graph. Figure 2 shows the volume of message for WhatsApp public groups.
4. Results and Discussion

4.1 Results from Sentiment analysis
In sentiment analysis, the author tries to find the people’s emotions in groups. Therefore, the author uses five classifiers to analyze the sentiment and uses Ensemble Learning to aggregate the results from these five classifiers to get the final results. Then the author can use this model to do the sentiment analysis. For this analysis, the author inputs user’s chat message in groups and gets the sentiment of this user. Figure 3 shows one result from this model when the author inputs some words.

```
print(now){sentiment("I am happy")}
print(now){sentiment("This movie is utter junk. There were absolutely 0 pythons. I don't see what the point was at all. Horrible movie. 0/10")}
```

Figure 3 result from this model when the author inputs some words.

4.2 Results from popularity analysis
In popularity analysis, the author makes statistics for the volume of message in each group and makes a bar graph to show the five most popular groups. Figure 4 shows the popularity bar graph.

```
```

Figure 4 popularity bar graph.

The author can find the five most popular groups in different time. For example, in the begin of June, the author can obtain the five most popular groups are 916036211434-1498983672@g.us, 912328282154@g.us, 917021760295-1482606046@g.us, 91988588697-1432904260@g.us and 913226785120-1422851978@g.us.

5. Conclusion
In this article, the author analyzes WhatsApp public groups in sentiment aspect. In sentiment analysis,
the author tries to find the sentiment of users in these groups. The author uses five classifiers to analyze the sentiment and voting classifier to aggregate the results to increase the accuracy and efficiency. The five classifiers are Naive Bayes classifier, Linear SVC classifier, Logistic Regression classifier, MNB classifier and Bernoulli NB classifier. These classifiers use different methods to do the analysis. The author inputs the training dataset and test dataset into this model and lets the five classifiers get the learning models. Then the author uses these five classifiers and voting classifier to find whether the user’s chat message is positive or negative.

In popularity analysis, the author makes statistics for the volume of message in each group and makes a bar graph to show the five most popular groups.

By analyzing the data, the author can find predict the sentiment of users. Also, there are many other aspects such as popularity and Multimedia Content to be analyzed. Therefore, the author will also try to analyze the other aspects to find more information in the future.

References
[1] Venkatesan, Rajasekar; Meng Joo, Er (2016). "A novel progressive learning technique for multi-class classification". Neurocomputing. 207: 310–321.
[2] Rish, Irina (2001). An empirical study of the naive Bayes classifier (PDF). IJCAI Workshop on Empirical Methods in AI
[3] Menard, Scott (2002). Applied Logistic Regression Analysis. SAGE. p. 91
[4] L. Xu, A. Krzyzak, C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition", IEEE Transactions on Systems Man and Cybernetics, vol. 22, no. 3, pp. 418-435, 1992.
[5] Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 35: 1-38