Public Opinion Analysis of Big Data Based on Machine Learning

Xuegang Chen, Hongbin Fan*, Sheng Duan, Luda Wang and Jialu Zhang
College of software and communication engineering, XiangNan University, Chenzhou,423000, China
*Corresponding Author: Hongbin Fan

Abstract: Network public opinions can be regarded as a collection of network comments, that is, a network comment can be regarded as the basic unit of network public opinion, based on this, network comment can be divided into spam comment, subjective comment and objective comment, we combine feature extraction method with machine learning method, machine learning method includes naive bayes and threshold delineation, and classify different types of network comments. By a large number of experiments, we have proved that the classifying accuracy of network comments is high, and the machine learning method can reduce the noise of text classification.

1. Introduction
Some experts[1] believe that public opinion is a collection of public sentiments, that is, in the face of a large amount of information dissemination on the internet, the public shows their different opinions on their concerns, which can be translated into public attitudes towards different events. Network public opinion has many characteristics[2], which include the freedom, real-time, multi-channel, anonymity, personalization and so on. Internet public opinion can be disseminated by e-mail, forum, blog and other means.

Network comment can be regarded as text, the tendency analysis of text[3], they can express users'opinions on different events, these opinions or comments can be further divided into negative comments and positive comments, by the tendency analysis of network public opinion, we can understand the emotional changes of people in the course of event's development. Tendency analysis of online comments is essentially an emotional analysis, which includes unsupervised[4] and supervised methods[5]. In part of the research[6], pattern recognition is introduced to describe the characteristics of reviews as a pattern, and the reviews are classified as objects. Some studies[7] did corpus analysis of network comment, sorted out different corpus and generated corpus. Through this way, vocabulary in different fields can be correlated, and better information classification conclusion can be obtained. As a result, some scholars[8] analyze the weights of lexicon of corpus, by this way, vocabulary of different fields can be correlated, and better information classification results can be obtained. Some scholars analyzed the weight of lexicon in corpus, the weight of lexicon lacks appropriate scenarios, so the weight is inaccurate. Some scholars[9] put corpus into classifying comments of Twitter, they confirmed that corpus can achieve better results in the face of simple sentences, but in the face of complex sentences, there are a lot of errors in the analysis of lexicon, which indicates that the amount of data in the corpus is limited, so the corpus cannot distinguish complex sentences.

People put machine learning into classifying network public opinion, and many researchers adopted this way to analyze emotion of different sentences. some researchers[10] used the method of
calculating the topic probability, we extract the Topics of different blogs, and divide the blogs related to the topic into one category. Nair, D. S et al[11] took movie reviews as an object of study, this paper establishes a topic classification model of reviews by using support vector machine. At the same time, it introduces corpus as a basis, which effectively improves the reliability of the classification of reviews.

In this paper, we use machine learning method to analyze public opinion. Network public opinion can be regarded as a collection of network comments, that is, a network comment can be regarded as a record of public opinion analysis. According to the characteristics of the records, we classify the comment as the category of spam comment, the category of subjective comment and the category of objective comment. We combine feature extraction method with classification algorithm, which includes decision tree algorithm and the logistic regression method. These classification algorithms can classify network comments according to certain constraints. Through statistical analysis of a large number of experiments, different types of comments are classified into the categories of junk comments, subjective comments and objective comments, and the classification results are further tested.

2. Public Opinion Analysis Technology Based on Machine Learning

2.1. Supervised Learning Algorithm

(1) Naive bayesian algorithm, the principle of the algorithm is relatively simple, different data categories \{a_1, a_2, a_3...\} as a possible classification feature, the proportion of different data classes/datasets is calculated to get a larger proportion of classes. Thus, the larger proportion of classification features is defined as corresponding \( a_i \). The implementation of the algorithm is based on a basic assumption, different data categories or different data features are independent of each other, therefore, it can reduce the computational complexity of the algorithm, at the same time, in real life, different data categories can not meet the assumption, but from the perspective of classification, the algorithm is very practical.

(2) Support Vector Machine (SVM) algorithm, which is based on the minimization of structural risk, it can be used to identify small samples, classify non-linear samples, relax variables and kernels as constraints, and transform non-linear data into linear data. At the same time, kernels can effectively reduce the computational complexity of data.

(3) K-distance aggregation algorithm, which takes the vector distance between data as the basis of judgment, randomly chooses K data objects as the initial category center, and corresponding data objects contain multiple features, that is, each data object can be regarded as a specific category. The judgment formula of distance, generally, the angle cosine theorem is put into better calculating the similarity of different data objects, the calculation method of euclidean distance can also be put into calculating the similarity of different objects, but the amount of calculation is too large, therefore, it is appropriate to put the angle cosine theorem into calculating the distance. We adopt K-distance aggregation algorithm, K value needs to be adjusted according to the type of data, the initial stage of calculation can be designated by ourselves, according to the effect of classification, constantly update K value.

2.2. Active Learning Algorithm

Artificial learning algorithm needs to extract part of the samples for testing. At the same time, the sampling is not for easily determined samples, but for uncertain samples. Data can be classified based on certain characteristics. In some cases, the data can be classified into category A or B, which means the classification of data can not be judged. Assuming that there are only two possibilities for a classification, if a sample belongs to class A, the probability reaches 80% and belonging to class B, the probability is 20%, then the class of the sample can be determined as class A, because the probability of the sample belonging to class A is higher. If the probability of a data item belonging to Class C is much less than that belonging to Class D, it is determined that the data item belongs to Class D. If a sample belongs to class A, the probability is 45%, and belonging to class B, the probability is 55%, then the class of the sample can be determined as class A or class B, there is uncertainty in the
determination of the sample, therefore, the sample needs to be adopted active learning algorithm to classify.

2.3. Semi-supervised Learning Algorithm
(1) Joint training algorithm includes collaborative steps and training steps. This algorithm adds one step, that is, training the model, selecting and marking a limited number of data sets. Each data set, data set and a selected algorithm are taken as a whole to correct the model, so as to determine the corresponding different data sets. Classification model. 3 classifiers belong to the same classification algorithm, but the parameters of these classifiers are different, normally, data objects to be classified will get the same classification results in 3 classifiers. In the actual classification process, the same data object may get different classification results in different classifiers, the classification results need to be counted, and the classification results with high frequency are determined as the classification results of data objects.

(2) Label propagation algorithm, which assumes that there is a certain relationship between data objects, each data object can be seen as the fusion of multiple attributes, there is a directional relationship between attributes and attributes, different data objects decompose the set of multiple attributes, each attribute can be seen as an entity, and the entity and entity can use directional relationship. Line segment representation, thus, there may be loops between attributes and attributes, and the relationship between attributes can be described by a graph data structure within different data objects. In the process of execution, the data object contains a complete data structure, and the data object with classification marks can be regarded as a classification feature of data. The links within the data object can describe the directional relations of different data objects. The more directional relations there are, the closer the relationship between data objects will be. The data object can be regarded as a reference object. The pointing relationship within the data object can help the known data object find the nearby data object. The more connection between the data object and the data object, the larger clusters between the data and the data can be constructed. Label propagation algorithm regards data objects as a set of attributes, and classifies data objects in a fine-grained manner. Texts can be composed of phrases. The same vocabulary within the text can establish links between different texts, which can be used for text correlation analysis.

3. Classification of Public Opinion Information
Public opinion information appears in the form of network comments, but network comments include spam comments, subjective comments and objective comments. The classification of these comments needs to extract the user's feelings. Subjective comments involve a large number of user ideas, rather than subjective comments containing incorrect opinions and lack of emotional statements. It is meaningful to get subjective comments from a large number of online comments.

3.1. Characteristics and Classification of Spam Network Comments
The expression forms of spam comments are constantly changing, there is no clear form for such comments, the meaning they express is not related to the theme, that is, the meaning they express has nothing to do with the theme. Spam comments can be divided into two categories, one is not related to the theme, for example, insulting others, language contains unhealthy information. Another is to regard pictures and links as the main body of comments and publish them in the form of advertisements. The common characteristics of spam comments are, they have short sentences, serious subjectivity, non-standard language, put pictures and links into inducing propaganda.

Spam comments are abundant in forms of expression, and different features have different effects on comments, spam comments magnify the characteristics of some comments, which can induce people's thinking, as shown in Table 1.
Table 1. Examples of garbage-related weibo comment

| Examples                                      | Comment's category |
|-----------------------------------------------|--------------------|
| forward weibo@CNN @luhan @DonaldTrump        | garbage            |
| find a house, find a job, and come to http://wh.ganji.com. | garbage            |
| I always think that after 18, I will be 19, and after 19, I will back to 18. | normal comment |
| very disappointed with General Motors and their CEO, Mary Barra, for closing in Ohio, Michigan and Maryland. | normal comment |
| tariffs are working far better than anyone ever anticipated. | normal comment |

In Table 1, we can see that normal comments are complete sentences, while in spam comments, we can see a lot of links, such as "forward weibo @CNN @luhan @DonaldTrump". We can see links of CNN and links of Luhan on Weibo, such as "find a house, find a job, and come to http://wh.ganji.com". We clearly see that the links to the website appear in the comments in the form of statements. In spam comments, many comments are too short, such as "find a house, find a job..." without a subject, but the meaning is clear, in normal reviews, such as "I always think that after 18, I will be 19, and after 19, I will back to 18.", this sentence shows the young mentality of users. We need to read the whole sentence to understand its meaning.

3.2. Characteristics and Classification of Subjective Comments-objective Comments

We analyze and filter the characteristics of spam comments, which need to be further divided into subjective comments and objective comments, managers of different websites can grasp users' views on social events and current affairs of the country by analyzing subjective reviews, therefore, we need to further classify the remaining network comments.

Subjective comments emphasize the user's own ideas, users describe different social events and current affairs of the country, this description does not refer to the existing facts, but expresses by their own cognition. Everyone has his own cognition, everyone has abundant scientific knowledge and accumulates certain social experience, but the occurrence of social events and occurrence of current affairs of the country have characteristic of suddenness, which contains unknown factors, people need to sort out their own thinking, put their own thoughts, knowledge they have learned, the years' experiences they have gone, into searching for a answer of their own, so that they can get their own ideas. From another point of view, subjective comments are a kind of speculative comments, which need to be guessed, they are not accurate, but they can reflect most people's cognition.

Objectively comments, they are accepted theories and they represent the truth accepted by all human beings. Commentators will start from the event itself, analyze the fact that the event exists, and objectively describe the event, but this description cannot express the commentator's own views. Taking the characteristics of subjective/objective comments into consider, we divide them into subjective comments and objective comments, their examples are shown in Table 2.

Table 2. Examples of subjective/objective weibo comment

| Examples                                      | Comment's category |
|-----------------------------------------------|--------------------|
| the girls in textile university are so beautiful. | subjective       |
| because of the hard work of ordinary people, so society can be stable. | subjective       |
| the earth is round.                           | objective         |
| stan lee, as the father of marvel, start to describe the superhero dream of ordinary people from his youth. | objective       |
| General Motors is the largest automobile manufacturer | objective       |
in the United States.

In Table 2, we can see that the description of subjective comments has many modifiers, such as "so", "because of", the meaning of the sentence contains a strong personal color, such as "the girls in textile university are so beautiful.", the sentence describes the female students of textile university, the commentators describe the female students of Textile University as "so beautiful", but people can adopt different aesthetic standards of beauty, that is, to change a person to comment, he may describe the textile university girls are friendly. An objective comment is adopted to convey objective facts, such as "the earth is round.", this sentence refers to the shape of the earth, it is a fact recognized by a large number of people, in other words, if someone says that the earth is square, most people will make it clear that the statement is false.

3.3. Recognition of Subjective and Objective Comments Based on Machine Learning

Subjective/objective comments need feature extraction, at the same time, subjective and objective comments have different characteristics, subjective/objective comments can be described as feature sets, which are represented by vectors, they are recognized by machine learning method.

We identify subjective comments and objective comments, firstly, need to collect subjective/objective comments and extract the characteristics of subjective/objective reviews and express the characteristics of different types of comments in the form of vector space. Adopt naive bayesian algorithm to train the data model of vector sets. Select the sample data for performance testing. The specific flow of subjective/objective comment identification is shown in Figure 1.

![Figure 1. Identifying process of subjective/objective comments](image)

In Figure 1, subjective/objective comments need to judge the opposite words, for example, objective comments indicate objective faces, there are many special words in subjective comments' descriptions, for example, the degree of modifying events can use "very", "so", "so much", for example, modifying personal feelings, we can use "feel", "think", "maybe", in a lot of sentences, we can see many special emotional symbols, such as "!" "?" "_\_\_"" _, different types of features need to be established corpus.

Subjective comment can describe the commentator's personal feelings, we adopt 3 indicators to identify subjective comment and objective comment, they are recall rate \( P \), precision rate \( R \) and \( Tr \) value measurement, their calculation methods are as follows:

\[
P = \frac{L_1}{L_1 + L_2} \times 100\% \tag{1}
\]

\[
R = \frac{L_1}{L_1 + L_3} \times 100\% \tag{2}
\]
2 \times PRTR \quad (3)

In formula (1) (2) (3), $L_1$ indicates the number of comments that the classification result is the selected category and actual classification result is the selected category. $L_2$ indicates the number of comments that the classification result is the selected category but actual classification result is not the selected category. $L_3$ indicates that the number of comments that the classification result is not the selected category but actual classification result is the selected category.

4. Experimental Results and Analysis

We get a large number of network comments from different websites, among them, there are obvious differences between spam comments and normal comments, we choose different topics, which include "China's economy" and "China-US relation" and so on. In the training data set, there are 2487 000 spam comments, 2356 000 normal comments, a total of 4843 000. In the test data set, and there are 8480 000 spam comments and 9880 000 normal comments, a total of 1836 000. The statistical analysis of network comments is completed under the distributed computing framework MapReduce, as shown in Figure 2.

![Figure 2](image)

**Figure 2.** The execution flow of MapReduce in distinguishing spam comments/normal comments

In Figure 2, we put the collected network comments into MapReduce framework for calculation, in Map stage, network comments are divided based on different characteristics, in Reduce stage, the key of spam comments is 1, and the key of normal comments is 2. We analyzed the frequency of "1" and "2" based on different characteristics.

We analyze the characteristics of spam comments in the training data sets and put the parameter of word-frequency-difference into calculating the discriminative ability of different features, the results are shown in Table 3.

| Feature          | Garbage comment | Normal comment | Weight |
|------------------|-----------------|----------------|--------|
| text similarity  | 1585000         | 8470000        | 0.278  |
| sentiment lexicon| 7380000         | 2450000        | 0.193  |
| hyperlink        | 9570000         | 720000         | 0.354  |
| number of nouns  | 10110000        | 4680000        | 0.208  |
| length of sentence| 14390000       | 11680000       | 0.082  |
In Table 3, we can see that replying comments, spam words and links can distinguish normal comments from spam comments, among them, spam words involve abusive words, degrading language, violence or pornographic information. Links are expressed in the form of letters "+"/", without much subjective information, instead, they recommend links to a website. Reply comments are usually repeated information, and the length of vocabulary combinations are significantly short. We can normalize different features and corresponding weight, as shown in Table 4.

| Feature of garbage comment | Weight |
|---------------------------|--------|
| text similarity           | 0.133  |
| sentiment lexicon         | 0.092  |
| hyperlink                 | 0.169  |
| number of nouns           | 0.099  |
| length of sentence        | 0.039  |
| spam words                | 0.225  |
| reply                     | 0.243  |

In Table 4, we normalize the weights of different features, in order to distinguish normal comment/spam comment, we can take replying comment and spam vocabulary as 2 main features, while affective dictionary, number of nouns and the length of sentence can not be the main features, because their weights are too small.

After eliminating the spam comments, we select some comments as the training set of subjective/objective comments, some comments as the testing set of subjective/objective comments, among them, in the training data set, there were 43000000 subjective comments and 41560000 objective comments, a total of 84560000. In the testing data set, there were 2169,000 subjective comments and 1575,0000 objective comments, a total of 374,400,00.

We extract the features of the remaining comments, denote the different feature sets by vector space, place them in bayesian classifier, and establish the recognition model of subjective/objective comment, the results are shown in Table 5 and Table 6.

| Categorizing result is subjective comment | Actual result is subjective comment | Actual result is objective comment |
|------------------------------------------|-------------------------------------|----------------------------------|
| categorizing result is subjective comment| 17340000                            | 6050000                          |
| categorizing result is objective comment | 4180000                             | 8370000                          |

| Testing data | Accuracy rate (%) | Recall rate (%) | Tr (%) |
|--------------|--------------------|-----------------|--------|
| subjective comment | 74.1%            | 80.7%           | 77.5%  |

In Table 5 and Table 6, we can see that the recognition rate of subjective comments is only 74.1%, this shows that some subjective sentences are difficult to distinguish, We distinguish the categories of subjective/objective comments, choosing features include the degree words, subjective mood words and special symbols of sentences, but in real life, many sentences contain many emotional words,
these words need to be analyzed as features, such as "Phelps swims in the pool like a shark paddles in the ocean.", in this sentence, "like" is also a word that modifies emotions, we need to choose different emotional words as the basis of division.

5. Conclusion
we analyze public opinion by doing a lot of experiments, adopting machine learning method, the recognition rate of subjective and objective comments needs to be improved, we need to select different emotional words as the basis of classification.

6. Acknowledgement
The authors would like to thank for financial support by youth fund project of the humanities and social sciences of Education Ministry (15YJC870004), science and technology innovation team of XiangNan University, Hunan province undergraduate research-based learning and innovative experimental project(719), social science planning project of Chenzhou(Czsskl2017067) and Institute of big data science of XiangNan university.

7. Reference
[1] Leeper, T. J. , & Slothuus, R. . (2014). Political parties, motivated reasoning, and public opinion formation. Political Psychology, 35(s1), 129-156.
[2] Ma, N. , & Liu, Y. . (2014). Superederank algorithm and its application in identifying opinion leader of online public opinion supernetwork. Expert Systems with Applications, 41(4), 1357-1368.
[3] Gui, B. , & Yang, X. . (2014). Research on Sentiment Tendency Analysis of Microtext Based on Sense Group. International Conference on Computational Intelligence & Security. IEEE.
[4] Siva, P. , Russell, C. , Xiang, T. , & Agapito, L. . (2013). Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection. IEEE Conference on Computer Vision & Pattern Recognition. IEEE.
[5] Kingma, D. P. , Rezende, D. J. , Mohamed, S. , & Welling, M. . (2014). Semi-supervised learning with deep generative models. Advances in Neural Information Processing Systems, 4, 3581-3589.
[6] Flanagan, D. P. , & Schneider, W. J. . (2016). Cross-battery assessment? xba psw? a case of mistaken identity: a commentary on kranzler and colleagues “classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth”. International Journal of School & Educational Psychology, 4.
[7] Rosenberg, A. E. . (2013). Who classification of soft tissue and bone, fourth edition: summary and commentary. Current Opinion in Oncology, 25(5), 571-573.
[8] Zaidan, O. F. , & Callisonburch, C. . (2013). Arabic dialect identification. Computational Linguistics, 40(1), 171-202.
[9] Bisschop, R. , Boomsma, M. F. , Van Raay, J. J. , Tiebosch, A. T. , Maas, M. , & Gerritsma, C. L. . (2013). High prevalence of pseudotumors in patients with a birmingham hip resurfacing prosthesis: a prospective cohort study of one hundred and twenty-nine patients. Journal of Bone & Joint Surgery-american Volume, 95(17), 1554-1560.
[10] Katchy, A. , Pinto, C. , Jonsson, P. , Nguyenvu, T. , Pandelova, M. , & Riu, A. , et al. (2014). Coexposure to phytoestrogens and bisphenol a mimics estrogenic effects in an additive manner. Toxicological Sciences, 138(1), 21-35.
[11] Nair, D. S. , Jayan, J. P. , Rajeev, R. R. , & Sherly, E. . (2015). Sentiment Analysis of Malayalam film review using machine learning techniques. International Conference on Advances in Computing. IEEE.