Research on Detection Method of Unhealthy Message in Social Network

Yabin Xu[1,2(✉)], Yongqing Jiao[2], Shujuan Chen[2], and Yangyang Li[3]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing, China

[2] Beijing Information Science and Technology University,
Beijing 100101, China
xyb@bistu.edu.cn

[3] China Academy of Electronics and Information Technology,
Beijing 100041, China

Abstract. In order to avoid the release and dissemination of eroticism, gamble, drug and politically sensitive message in social network, and purify the network space, we propose a method to detect unhealthy message in social network. Firstly, the Naive Bayes model is used to classify the message released by the social network. Then, according to the features of all kinds of unhealthy message, the classification model of Support Vector Machine (SVM) is used to make further judgment. The comparative experiment results show that, the classification model of SVM has better precognitive effect than that of Naive Bayes and Decision Tree.

Keywords: Social network · Unhealthy message detection · Naive Bayes · Support vector machine

1 Introduction

With the emergent and rapid development of social network, people can communicate and comment without geographical constraints, conveniently and fast. According to the 39th China Internet Development Statistics Report, as of December 2016, the usage rate of WeChat friends circle and Weibo were respectively 85.8% and 37.1%.

Due to the features of convenience and fast and wide spread, the social network has become a tool for criminals to disseminate unhealthy messages about eroticism, gamble, drug and also politically sensitive message through social network. The spread of unhealthy message not only endanger the physical and mental health of Internet users, especially young people [1], but also impact the security and stability of a country. So, curbing the release and dissemination of unhealthy and sensitive message in social networks have become the extremely urgent problem which needs to be solved.

Therefore, how to quickly and efficiently detect the unhealthy message in social network can not only protect the users from the effect of unhealthy message, but also promote the healthy development of social network effectively. But the solving method is still a combination of artificial and mechanical identification, which is mainly

X. Sun et al. (Eds.): ICAIS 2019, LNCS 11632, pp. 497–508, 2019.
https://doi.org/10.1007/978-3-030-24274-9_45

because the effect of mechanical identification is limited. Therefore, to improve the accuracy and efficiency of machine identification has a huge development space.

2 Related Work

At present, the major social network platforms take measure of matching sensitive words as the basic means to detect unhealthy messages. However, the unhealthy message releasers will deform the sensitive words to avoid being detected. Therefore, the detection of pornographic, gamble, drugs and even politically sensitive content in social network only by detecting sensitive words is not enough.

Cohen [2] proposed a rule learning method for mail classification. He analyzed the particularity of e-mail messages and presented a new approach to learning sensitive words based on RIPER rules. Salvador Nieto Sanchez et al. of Louisiana State University use the OCAT mining algorithm for text classification [3]. This method uses the Boolean model to break down the text and ignores the weight of the word.

Wang [4] and others proposed a Naive Bayesian classification model based on five kinds of detection feature. But most features do not have robustness. Zhang [5] firstly used the regular expression to deal with special words or symbols, and then constructed a filtration model of unhealthy text in virtual community by using the algorithm of Bayesian and support vector machine. Fang [6] employed a multi-label classification framework and the method have better accuracy and lower FNR. Xiao [7] proposed a SMK-means which is achieved by Mini Batch K-means based on simulated annealing algorithm for anomalous detection of abnormal information.

Zhu [8] researched text filtration method based on synonym extension, and found synonyms through thesauruses and Internet mining for query expansion. Zhou [9] extracted the set of positive and negative features that can represent the content of the positive and negative text, and designed a new text weight calculation method.

Jinghong [10] analyzed the problem of pornographic message in social network. Huiyu [11] used a new neural network algorithm to design a real-time monitoring prototype system for unhealthy messages. Shao [12] proposed a kind of information filtration method based on semantic association. By calculating the relevant degree of meaning between the text and unhealthy semantic factors, which effectively overcome the drawbacks of traditional methods.

Meng [13] analyzed the application of semantic analysis technology in the field of counter-terrorism and showed the importance of semantics for sensitive content analysis. Liu et al. [14] adopted a sensitive content filtering approach based on a two-stage filtering model (topic information filtering and propensity filtering). In recent years, with the development of deep learning, Neerbeky [15] used recurrent neural networks to assign sensitivity scores to the semantic components of the sentence structure, enabling interactive detection of sensitive information.

From the above research, it can be seen that the current academic circles have carried out a number of research on the detection of unhealthy message on Internet, but the research on the detection of unhealthy message on social network is numbered. In addition, because the social network has their own characteristics, such as short text and colloquial. The detection method is different from the web page with the characteristics of long text, logical, written language, thus the method is more difficult.

In this paper, the method of identifying unhealthy message in social network is as follows. In the first step, we should preprocess the data by the methods of word segmentation, disambiguation, dereference, and so on, we use the Naive Bayesian method to classify the message of social network into four categories which is highly related to eroticism, gamble, drug and politically sensitive and other categories. We only care about the first four categories. In the second step, based on the feature extraction from unhealthy message in each category, the support vector machine (SVM) model determines if it is unhealthy message. If a certain type of unhealthy message is determined, the release of the message is rejected.

3 Classification of Released Message

This paper uses the naive Bayes method to classify the released content. The method is efficient, accurate and easy to implement [16]. The specific processing is as follows:

$C = \{c_1, c_2, \ldots, c_5\}$ denotes the collection of categories, $c_1 \sim c_4$ respectively stand eroticism, gamble, drug and politically sensitive categories, c_5 represents other categories. Each released content is represented by vector $d=\{t_1, t_2, \ldots, t_n\}$. We use D to denote the example set that contains all released contents, so the relationship is $d \in D$.

$p(c_i|d)$ represents the probability that the released content d belongs to category c_i. If the probability value of $p(c_i|d)$ is bigger, the released content is more likely belongs to the category c_i. According to Bayes formula, the probability that released content d belonging to category c_i is:

$$p(c_i|d) = \frac{p(d|c_i)p(c_i)}{p(d)} = \frac{p(d|c_i)p(c_i)}{\sum_{j=1}^{5} p(d|c_j)p(c_j)} \tag{1}$$

$P(c_i)$ is the probability that the category of released content is c_i. We suppose that $|D_c|$ is the number of all released contents, and $|D_{c_i}|$ is the number of the released content of category c_i. So we can get the following equation.

$$p(c_i) = \frac{1 + |D_{c_i}|}{|C| + |D_c|} \tag{2}$$

$p(d|c_i)$ can be calculated by formula (3):

$$p(d|c_i) = p((t_1, t_2, \ldots, t_n)|c_i) = \prod_{j=1}^{n} p(t_j|c_i) \tag{3}$$

Among them, $p(t_j|c_i)$ indicates the probability of occurrence of the word t_j in category c_i. According to the Multi-variable Bernoulli Model, among $d = \{t_1, t_2, \ldots, t_n\}$, $t_k \in (0, 1)$, $1 \le k \le n$, $t_k = 0$ means that t_k does not appear in the vector d. $t_k = 1$ means that t_k appears in the vector d. $|D_{c_{ik}}|$ denotes the number of released content that contains the word t_k and belongs to category c_i. In order to prevent the occurrence of zero probability, we can add a smoothing factor:

$$p_{ik} = p(t_k = 1|c_i) = \frac{|D_{c_{ik}}| + 1}{|D_{c_i}| + 2} \tag{4}$$

So we can get the following equation:

$$p(t_k|c_i) = p_{ik}^{t_k}(1 - p_{ik})^{1-t_k} = \left(\frac{p_{ik}}{1 - p_{ik}}\right)^{t_k}(1 - p_{ik}) \tag{5}$$

According to the formula (2)–(6), we can get the following

$$\begin{aligned} \hat{c}(d) &= arg \max_{c_i \in C}\left\{p(c_i) \times \prod_{k=1}^{n}\left(\frac{p_{ik}}{1 - p_{ik}}\right)^{t_k}(1 - p_{ik})\right\} \\ &= arg \max_{c_i \in C}\left\{log\left(\frac{1 + |D_{c_i}|}{|C| + |D_c|}\right) + \sum_{k=1}^{n} log(1 - p_{ik}) + \sum_{k=1}^{n} t_k \times log\left(\frac{p_{ik}}{1 - p_{ik}}\right)\right\} \end{aligned} \tag{6}$$

From Eq. (6), it can be seen that the words with $t_k = 1$ play a role in the classification. In other words, we can get the category c_i from the released content d, by seeking i to make the value of $\hat{c}(d)$maximum.

The steps of content classification are as follows:

(1) We first take a piece of data from Content Category table, segment words, remove stop words, so we obtain $d = \{t_1, t_2, \ldots, t_k, \ldots, t_n\}$. Then we write ($t_k$, Count, Category, Privacy Category) in the Word Category table, where the initial value of Count is 1. If (t_k, Category, Privacy Category) is already in Word Category table, then we plus 1;

(2) We add the count of the corresponding category in the Category Table with 1;

(3) Repeat step (1), (2);

(4) Take a piece of data from the example set, preprocess and remove stop words and repeat words, so $d = \{t_1, t_2, \ldots, t_k, \ldots, t_n\}$. is obtained;

(5) Calculate the value of (2), (4), (6), and get i which makes $\hat{c}(d)$ maximum. Then c_i is the category of released content d.

4 Feature Recognition of Unhealthy Message

In this paper, after classifying all the content from the data set, we give a detailed feature identification process by taking pornographic category as an example.

4.1 Camouflage Feature Recognition of Sensitive Words

To avoid sensitive words being detected, most of the words in unhealthy message are disguised with the following characteristics. We need to adopt the corresponding processing method respectively:

(1) Recognition method of sensitive words with special symbols.

In unhealthy messages, criminals often separate sensitive words with special symbols to outsmart the detection. In order to determine the legitimacy of released messages, we can adopt the method of regular expression to match special symbols such as "*", "&" and others and remove them. Then we restore the sensitive words into the state of the normal combination. For example: for "practice 法*轮*功, you can keep fit", we use regular expressions to identify "*" and delete it. Then, we can get the sensitive word "法轮功" without word segmentation.

(2) Recognition method of sensitive words with pinyin instead of word.

In the released text with unhealthy content, sensitive words are often partly replaced by phonetic symbols, such as: "法lun功". For this case, literature [17] proposed a method to find the most similar word which can combine with adjacent words and establish a comparison table of phonetic alphabet and word. We accumulate the number of occurrence of the word in the text in order to count the frequency of sensitive words in documents. The words can be quickly matched with a phonetic alphabet. Therefore, we use this method to identify sensitive word in the pinyin instead of the word.

(3) Sensitive word recognition method of components.

Criminals often split words in sensitive words to avoid detection. For example, the "轮" in "法轮功" is splited into "车仑" to avoid detection. The literature [18] proposed a method, in which we firstly find components of Chinese characters appeared in released content, and judge whether the word adjacent to the right is a component of Chinese characters. If the word exists in the dictionary, the ability of combining the word with its adjacent word to form a sensitive word is found by using the word segmentation algorithm. If the word doesn't exist in the dictionary, we go to identify the next word. This article learn from this method.

4.2 Similarity Feature Recognition of Released Message

Unhealthy messages transmitted through such social network may be overwritten in a short period of time. They repeatedly release unhealthy message at different time to ensure that unhealthy messages are disseminated. So, we can take the mean of similarity of messages released by user as a feature of the unhealthy message.

We firstly delete url (links), @ (symbols), emotion faces, etc. Next, we segment words and remove the stop word. Finally, we calculate the similarity of released messages of a particular user in adjacent time, according to the TF-IDF algorithm and the cosine similarity. So we can take the mean of similarity value of all released content as a feature of unhealthy messages.

In the TF-IDF method, the term frequency (tf) refers to the frequency that a word appears in a given released message:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{k=m} n_{k,j}} \tag{7}$$

In Eq. (7), $n_{i,j}$ indicates the number of times word t_i appears in the released message d_j.

The inverse document frequency (idf) refers to the universal importance of a word, that is, the idf of a particular word is obtained through dividing the number of released messages by the number of released messages containing the word:

$$idf_i = log\frac{|D|}{1 + \left|\left\{j : t_i \in d_j\right\}\right|} \tag{8}$$

In Eq. (8), |D| indicates the total number of released message, $\left|\left\{j{:}t_i \in d_j\right\}\right|$ indicates the number of released messages containing the word t_i.

The $\mathrm{tfidf}_{i,j} = tf_{i,j} \times idf_i$ value of each feature word in the released message is used as the weight of the word $w_{i,j}$ to construct the vector space model, the j-th released message d_j is expressed as: $d_j = \left\{w_{1,j}; w_{2,j}; \ldots, w_{n,j}\right\}$.

In the vector space model, the cosine of the angle between the vectors is used to represent the message similarity of the two texts:

$$sim\left(d_i, d_j\right) = cos\theta = \frac{\sum_{k=1}^{n} w_{k,i} \times w_{k,j}}{\sqrt{\left(\sum_{k=1}^{n} w_{k,i}^2\right)\left(\sum_{k=1}^{n} w_{k,j}^2\right)}} \tag{9}$$

Then, we take the average value of similarity as the similarity of user release message:

$$Avg_{sim} = \frac{\sum_{i=1}^{n-1} sim(d_i, d_{i+1})}{total_{content} - 1} \tag{10}$$

4.3 Content Feature Recognition of Unhealthy Message

The Proportion of URL Links in Released Message. The outlaws usually release eye-catching keywords, add links to arouse the curiosity of ordinary users, and entice users to click the link so as to disseminate unhealthy messages. Therefore, the ratio of url link can also be used as one of the features of unhealthy messages.

We use U to indicate the percentage of the released messages containing url links to the total number of messages released by users.

The Proportion of Hot Topics in Released Messages. Lawless people always make use of some current hot topics of the time, and add some hot topics to their unhealthy message. This article uses H to represent the proportion of released messages containing hot topics to the total number of released messages.

The Number of @ in Released Messages. Lawless people often use this method of @ to push unhealthy messages directly to the concerned users. In this paper, we use M to express the mean of each released message included symbol @.

The Proportion of Picture Included in Released Messages. Criminals use very eye-catching pictures to attract normal users to click and link to an unhealthy information page. In this paper, we use P to indicate the mean of the picture number included in each message.

4.4 Feature Recognition of Unhealthy Message Releaser

Concern Degree of Unhealthy Message. The fan number of lawless people is usually less, and the concern degree tends to be low because the content is often involved in unhealthy message. In this paper, we use A to express the attention degree, that is, the proportion of the number of messages concerned by their fans to the total number of messages released by the user.

The Ratio of Numbers of Concerned Users to the Numbers of Followers. As garbage users often release unhealthy message, the number of fans is poor. But garbage users will concern a large number of users in order to achieve the purpose of expanding the scope of unhealthy messages. Therefore, the number of span users watching is far greater than the number of followers. In this paper, we use F to indicate the proportion of the number of concerned users to the number of fans.

5 Determination of Each Class of Unhealthy Message

In this paper, we need to identify the characteristics of the bad information in the four types of information to further determine whether it is bad information. We only take the eroticism class as example. According to the features determined in Sect. 4, we determine whether the content about pornographic is unhealthy message by using the support vector machine model.

According to the need of classification, this paper takes the classification of two kinds of data as an example. Given the training set $(\mathbf{x_i}, \mathrm{y_i}), \mathrm{i} = 1,2,\ldots,\mathrm{l}, \mathbf{x_i} \in \mathrm{R^n}, \mathbf{x_i}$ represents the characteristic vector of the first i sample; $\mathrm{y} \in \{\pm 1\}$, represents the category of the i-th sample hyperplane: $(\mathbf{w} \cdot \mathbf{x}) + \mathrm{b} = 0$. In order to classify all samples in the data set correctly and possess the classification interval, it is necessary to satisfy the constraints: $\mathrm{y_i}[(\mathbf{w} \cdot \mathbf{x_i}) + \mathrm{b}] \geq 1, \mathrm{i} = 1,2,\ldots,\mathrm{l}$. Classification interval is $2/||\mathbf{w}||$. So, the problem of constructing the optimal hyperplane is converted to the constraint condition (11):

$$min\emptyset(w) = \frac{1}{2}||w||^2 = \frac{1}{2}(w' \cdot w) \tag{11}$$

We introduce the Lagrange function to solve the constrained optimization problem. The problem of constrained optimization is determined by the saddle point of the Lagrange function, and the solution of the optimization problem is that the bias of the w and b at the saddle point is equal to 0. Therefore, we transform QP (quadratic programming) problem into following dual problem:

$$maxQ(a) = \sum_{j=1}^{l} a_j - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \left(x_i \cdot x_j\right) \qquad s.t. \sum_{j=1}^{l} a_j y_j = 0, \quad a_j \geq 0, j = 1,2,\ldots,l. \tag{12}$$

The optimal solution can be obtained as: $\mathbf{a}^* = \left(\mathrm{a}_1^*, \mathrm{a}_2^*, \ldots, \mathrm{a}_1^*\right)^{\mathrm{T}}$.

The optimal weight vector $\boldsymbol{w}^*$ and the optimal bias $\boldsymbol{b}^*$ are:

$$\boldsymbol{w}^* = \sum_{j=1}^{l} a_j^* y_j x_j \qquad \boldsymbol{b}^* = y_i - \sum_{j=1}^{l} y_j a_j^* \left(x_j \cdot x_i\right) \tag{13}$$

In the above formula, the subscript $\mathrm{j} \in \left\{\mathrm{j}|\mathrm{a}_\mathrm{j}^* > 0\right\}$, we can get the optimal classification hyperplane $(\mathbf{w}^* \cdot \mathbf{x}) + \mathrm{b}^* = 0$, and the optimal classification function is:

$$f(x) = sgn\{(w^* \cdot x) + b^*\} = sgn\left\{\left\{\sum_{j=1}^{1} a_j^* y_j \left(x_j \cdot x_i\right)\right\} + b^*\right\}, x \in R^n \tag{14}$$

The goal of training based on SVM is to obtain the classification surface that optimizes the structural risk. The training process is as follows (12) ~ (14).

6 Experiment

6.1 Data Set and Data Preprocessing

The data set used in this paper is Datatang. The data set consists of user table and Weibo table. The preprocessing of the data set is as follows: Firstly, we carry on the Chinese word segmentation, delete stop word and other processing on content field of Weibo Table. Finally, we obtain a total of 8283398 experimental data of 1017553 users. In this paper, we sort them out manually that stop word dictionary containing 1893 stop words, the unhealthy dictionary containing 317 words and the health feature dictionary containing 523 words.

6.2 Evaluation Indicator

In this paper, bi represents unhealthy message, N_{bi} represents the number of unhealthy message bi, $\overline{\mathrm{bi}}$ represents non-unhealthy message, which is normal message, $\mathrm{N}_{\mathrm{bi} \to \mathrm{bi}}$ represents the number of unhealthy message bi correctly detected, $\mathrm{N}_{\mathrm{bi} \to \overline{\mathrm{bi}}}$ represents the number of unhealthy message bi improperly detected, $N_{\overline{bi} \to bi}$ represents the number of error that $\overline{bi}$ is classified as bi, $N_{\overline{bi} \to \overline{bi}}$ represents the number of $\overline{bi}$ that is correctly detected. In this paper, the detection effect is measured by the recall rate, precision and comprehensive index F-Measure [19]. The three indicators are defined as follows (Table 1):

Table 1. Calculation method of each index.

Recall rate	Precision rate	F-Measure
$R = \frac{N_{bi \to bi}}{N_{bi}}$	$P = \frac{N_{bi \to bi}}{N_{bi \to bi} + N_{\overline{bi} \to bi}}$	$F = \frac{2 \times P \times R}{P + R}$

In addition, we introduce the ROC (Receiver Operating Characteristic curve) and AUC (Area Under roc Curve) in order to overcome the shortcomings of traditional evaluation indexes when measure the classifier under the condition that the positive and negative samples are not balanced. The ROC curve is measure by TPR (True Positive Rate) and FPR (False Positive Rate). TPR represents the probability of correctly detecting the positive case. FPR represents the probability that a negative example is divided into positive cases.

In the generated ROC space, the TPR is the vertical coordinate, and the FPR is the horizontal coordinate. The quality of the classifier can be measured directly with the AUC area, which is the area under the ROC curve. It can be seen from the ROC curve that the performance of the classification algorithm is better when the TPR grows rapidly, the ROC curve approaches the longitudinal axis quickly, and the area of AUC is large.

6.3 Experiment and Result Analysis

Classification Experiment on Released Messages. The data set used in this paper includes 8283398 data. After pretreatment and manual labeling respectively, we get 53998 piece of eroticism data, 47395 piece of gamble data, 51682 piece of drug data, 49257 piece of sensitive political data and 8081066 piece of data in other category. In this paper, naïve Bayes classifier in weka3.8.0 software package is used, and the experimental results are shown in Table 2:

Table 2. Experimental results of content classification.

Category	Recall	Precision	F-Measure
Eroticism	84.52%	83.10%	83.80%
Gamble	82.71%	81.61%	82.16%
Drug	84.11%	83.14%	83.62%
Political	83.77%	84.07%	83.92%
Others	80.94%	81.29%	10 point, italic

As we can see from Table 2, the recall rate of each category is more than 80%, the precision rate is above 81%, and F-Measure is above 81%. The results show that naive Bayes classifier has a good classification effect on multi-classification, and has little difference for different categories. It means that the performance is more stable.

Experiment on Unhealthy Message Judgement. (1) Experiment on balanced positive and negative samples.

The data set used in this article, has 54,000 data on pornography, where 10672 data include unhealthy message, 43328 data does not include unhealthy message, 5000 data are used as example set and other 5000 data are used as test set. This paper use SMO (SVM algorithm) classifier, the naïve Bayes classifier and the trees.J48 (C4.5 decision tree algorithm) classifier in weka3.8.0 software package to carry out experiment, and the experimental results are shown in Table 3:

Table 3. Experimental results of content classification.

Algorithm	Precision	Recall	F-Measure
SVM	82.7%	82.7%	82.7%
Naïve Bayes	72.3%	72.3%	72.2%
C4.5	78.5%	78.5%	78.5%

From the experimental result, we can see that the SVM algorithm is better than the naïve Bayes algorithm and the C4.5 algorithm. Therefore, the SVM algorithm selected in this paper is better than the other two algorithms in judging the unhealthy message in social network in the case of high dimension and small samples.

(2) Experiment on unbalanced positive and negative samples.

In unbalanced positive and negative samples experiment, 25000 data were used as example set and another 25000 data were used as test set. Among them, there are 5000 unhealthy messages and 20000 healthy messages. In this paper, SMO (SVM algorithm) classifier, naïve Bayes classifier, the trees.J48 (C4.5 decision tree algorithm) classifier in weka3.8.0 software package were carried out respectively, and we got respective ROC curve shown in Fig. 1:

In Fig. 1, curve (a) represents the C4.5 decision tree algorithm, and AUC = 0.851. Curve (b) represents the SVM algorithm, and AUC = 0.837. Curve (c) represents the Naïve Bayes algorithm, and AUC = 0.757. We can see that some of the points in the C4.5 decision tree algorithm and the SVM algorithm are coincident. The classification effect of C4.5 decision tree algorithm is better than that of support vector machine, although the difference is not great. The experimental results of support vector machine algorithm and C4.5 decision tree algorithm are much better than that of naive Bayesian algorithm. Also, the experimental results of support vector machine are much better than that of C4.5 decision tree algorithm in balanced experiment. So, it shows that the classification effect of the SVM algorithm is the best.

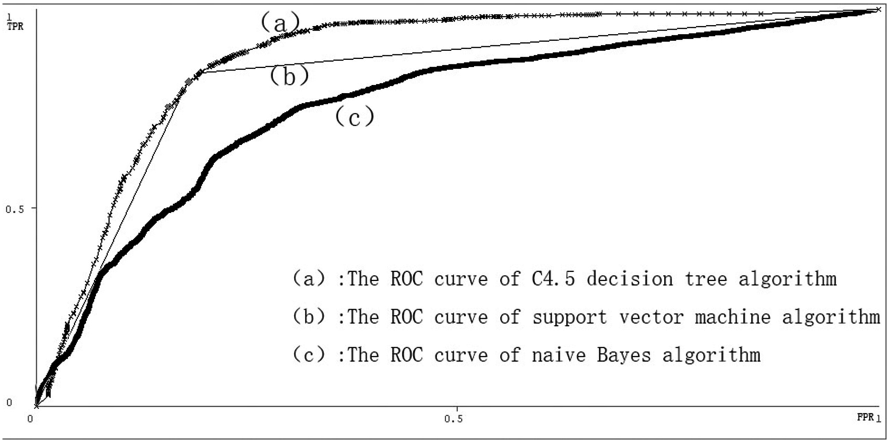

Fig. 1. The ROC curve.

7 Conclusion

The promotion and popularization of social network can effectively help people to establish and maintain the relationship with each other to meet the need of social contact. But there are many people with unhealthy habits to use social network to release unhealthy messages and illegal elements. They make use of the characteristics of social network to disseminate rapidly and widely in order to obtain unlawful gains or to achieve ulterior motives. So, in order to purify the cyberspace and suppress the spread of unhealthy message, the research on the detection of unhealthy message in social network is not only realistic but also far-reaching.

Based on the analysis of unhealthy messages of eroticism, gambling, drug and political sensitivity, this paper uses the Naive Bayesian method to perform multi-classification experiments on the characteristics of unhealthy message in social network. Then, based on the unique characteristics of various types of unhealthy message, the support vector machine algorithm, the C4.5 decision tree algorithm and the naive Bayesian algorithm are used in comparison experiments on balanced positive and negative samples and also unbalanced positive and negative samples. The experimental results show that the support vector machine algorithm is better than the other two classification algorithms.

Acknowledgement. This work was supported by the National Natural Science Foundation of China Nos. 61672101, the Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDDXN004)* and Key Lab of Information Network Security, Ministry of Public Security, No.C18601.

References

1. Wang, L.: 17-year-old high school students want to try rape girls after seeing porn pages [EB/OL], 10 October 2014. http://news.hsw.cn/s/2014/1010/162554.shtml
2. Cohen, W.W.: Learning rules that classify email. In: Proceedings of the AAAI Spring Symposium on Machine Learning & Information Access, vol. 96, no. 5, pp. 18–25 (2000)
3. Li, D.: Study of the information content security filter method in WEB, Shanxi University (2004)
4. Wang, A.: Don't follow me: spam detection in Twitter. In: Proceedings of the International Conference on Security and Cryptography, Athens, pp. 142–151 (2011)
5. Zhang, L.: Research on bad information filtering technology in virtual community, Kunming University of Science and Technology (2011)
6. Fang, S., et al.: Feature selection method based on class discriminative degree for intelligent medical diagnosis. CMC Comput. Mater. Continua **55**(3), 419–433 (2018)
7. Xiao, B., Wang, Z., Liu, Q., Liu, X.: SMK-means: an improved mini batch K-means algorithm based on mapreduce with big data. CMC Comput. Mater. Continua **56**(3), 365–379 (2018)
8. Zhu, X.: Research on adaptive text filtering system based on vector spatial model, Shandong Normal University (2006)
9. Zhou, J.: A bad text filtering method, University of Electronic Science and Technology of China (2016)
10. Jinghong, X., Yifei, L.: "Anti - pornography" in the social network age: status quo, problems and countermeasures. J. Beijing Univ. Posts Telecommun. (Soc. Sci. Ed.) **17**(3), 9–13 (2015)
11. Huiyu, H., Congdong, L., Jiadong, R.: Real Time Monitoring Prototype System for Bad Information Based on Artificial Neural Networks. Comput. Eng. **32**(2), 254–256 (2006)
12. Shao, X., Xu, Q.: Network camouflage bad information detection method of research and simulation. Comput. Simul. **29**(2), 135–138 (2012)
13. Meng, X., Zhou, X.P., Wu, S.Z.: The application research of semantic analysis in the field of anti-terrorism. J. Intell. **36**(3), 13–17 (2017)
14. Liu, M.Y., Huang, G.J.: Research on text filter model for information content security. J. Chin. Inf. Process. **31**(2), 126–131 (2017)
15. Neerbeky, J., Assentz, I., Dolog, P.: TABOO: detecting unstructured sensitive information using recursive neural networks. In: IEEE, International Conference on Data Engineering. IEEE (2017)
16. Jiang, Z.: Research on micro-blogging privacy detection based on Bayesian, Harbin Engineering University (2013)
17. Wei, S.: Harmful information detection and filtering toward interactive web media, Dalian Maritime University (2009)
18. Huiling, W., Xiwei, G., Jianjing, S., et al.: Research on the text pre-processing to malicious information filtering. Microcomput. Inf. **22**(12X), 58–60 (2006)
19. Yabin, J.X.: Privacy content detection method for the judgment documents. J. Chongqing Univ. Posts Telecommun. (Nat. Sci. Ed.) **27**(5), 639–646 (2015)