A Sensitive Words Filtering Model Based on Web Text Features

Rui Yao
Faculty of Information Technology,
Beijing University of Technology
Beijing, China
yaorui9501@163.com

Yang Cao
Faculty of Information Technology,
Beijing University of Technology
Beijing, China
caoyangcwz@emails.bjut.edu.cn

Zhiming Ding
Faculty of Information Technology,
Beijing University of Technology
Beijing, China
zmding@bjut.edu.cn

Limin Guo
Faculty of Information Technology,
Beijing University of Technology
Beijing, China
guolimin@bjut.edu.cn

ABSTRACT

The false advertising of food and drag on the Internet is mainly based on the content of the product website promotion pages. When people browse a website, they get the most parts of the information from texts on the web. In order to help people to distinguish whether it is false propaganda on this website, we propose a solution for identifying false advertising of text content on food and drug websites by designing the sensitive word recognition model. This paper introduces in detail the specific design and implementation of the food webpage text sensitive text recognition model, including the system improvement of text acquisition and word segmentation algorithm, feature extraction algorithm and text classification in the sensitive word list extraction. The detailed design and execution flow of the voting decision determination result algorithm of the five text classification algorithms are combined for filtering. Finally, we conducted a series of experiments, and the experimental results demonstrated that the proposed filtering solution is effective.

CCS Concepts

• **Theory of computation → Design and analysis of algorithms** • **Computing methodologies → Natural language processing** • **Security and privacy → Web application security**

Keywords

False advertisements; sensitive word discrimination; feature extraction; text classification; machine learning

1. INTRODUCTION

With the rapid development of the Internet, our daily life is surrounded by a variety of information. In the gradually expanding Internet drug transactions, criminals falsely publicize drug information and deceive consumers to purchase, which seriously damages people's interests [1]. How to help consumers to identify drug information intelligently with false propaganda through machine learning methods has become one of the important topics in the field of information filtering and data mining [2].

At present, the existing text-based sensitive word filtering is generally implemented by a combination of keyword list and text filtering or by using an algorithm such as a neural network to imitate the filtering information of the human brain [3,4]. The former customize the list of sensitive words according to the filtering requirements, and filters sensitive words in websites [5], forums or message comments; the latter requires the technical support of machine learning, according to the results of automatic classification filter information for a specific category. In this paper, we focus on the construction of text classifier in the sensitive word identification model [6]. The filtering of the false content on the website is to identify the category of the false and non-false propaganda of its text content.

In this paper, we proposed a sensitive word filtering model for false propaganda of food and drag websites. We designed our model data to conduct in our model, which includes text acquisition method, Chinese text segmentation, feature extraction and text automatic classification that is used in artificial intelligence. The main contributions of this paper are summarized as follows:

- A sensitive words filtering model is proposed in this paper for the fake publicity of the anti-drug website. The implementation solution provides a new idea for network content filtering and has certain application value.
- A novel feature matrix suitable for the text feature weight is designed to obtain the feature matrix of the corresponding text in our model.
- Regarding the textual judgment results of the model, based on the key search algorithm, the five filtering algorithms are combined, and the innovative method of voting for the final result is used to give a more reasonable and correct decision filtering result.

The rest of this paper is organized as follows. We first give a description of the sensitive word filtering model in Section 2. Section 3 presents our classifiers and voting strategy. We elaborate on the evaluation criteria and experimental results in Section 4. Finally, this paper is concluded in Section 5.

CSAI '18, December 8–10, 2018, Shenzhen, China
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6606-9/18/12…$15.00
DOI: https://doi.org/10.1145/3297156.3297232

2. OUR PROPOSED MODEL

In this paper, we propose a sensitive word filtering model for segmentation, feature extraction and classification of food and drug web pages, which helps consumers to identify the food and drug websites with false propaganda. At present, people get that information from the publication of relevant state agencies or news exposure. Our model provided one way to identify the false text of the website through machine learning automatically. Advocacy to achieve the purpose of filtering these sites, our model is mainly divided into three parts, the Chinese word segmentation, feature extraction and classifier construction, as shown in Figure 1.

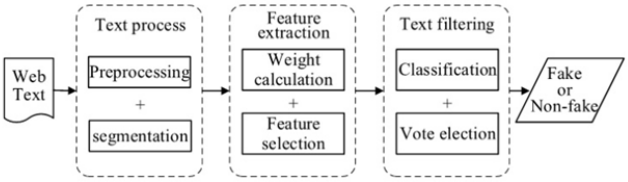

Figure 1. The system architecture of our model

2.1 Word Segmentation

Compared to standardized data, text content is structured, which is very limited (some or even no structure). Some have a certain structure, which focuses on the text format, rather than focusing on the text content, so the text data needs to be standardized (pre-processed). Stop words can generally be interpreted as some common and high-frequency words that often appear in various sentences.

The system segmentation method is mechanical word segmentation, using the maximum segmentation combination based on word frequency and the word segmentation method of the HMM model [7]. The process is as follows:

(1) Generating a word tree from a dictionary. (2) For these sentences, cut them into segments by the given dictionary. (3) After generating the word tree, the dictionary calculates the number of occurrences of each word feature in the constructed number of words and converts the number of times into the total frequency of occurrence of the word features. In dynamic programming, by locking the words that have been segmented in the sentence to be segmented, the frequency (number/total) of the feature of the word is obtained. If the word feature is not found, the word with the least frequency in the word tree is used. The frequency of the feature is used as the frequency of the feature, and the maximum probability path of the occurrence of a word feature is constructed, and then the segmentation combination of the maximum probability of the sentence to be segmented is obtained, which achieves an ideal effect. (4) For words that do not appear in the dictionary, according to the probability table and the Viterbi algorithm obtained from the previous training, a BEMS sequence with the highest probability can be obtained. With the HMM model, the Chinese vocabulary is marked according to the four states of BEMS. (5) Finally get the final cut combination.

2.2 Feature Extraction

After the word segmentation, we need to select the feature words that contribute to the text classification, and pick up some features with the best recognition or representativeness from the initial feature space. Traditional word frequency-inverse document frequency (TF-IDF)[8] tends to filter out common words and retain important words[9]. For example, words such as "and" appear more frequently in text, but cannot be used as feature words for classification. Our model is used to distinguish whether the text contains false propaganda. In the feature extraction process, it mainly includes two steps of weight calculation and feature extraction.

2.2.1 Weight calculation

In traditional TF×IDF tends to filter out common words and retain important words. The high word frequency within a particular file, and the low file frequency of the word in the entire file set, can produce a high weight TF-IDF.

$$\mathrm{W}\left(t,\vec{d}\right) = tf_{ij} \times idf_i = tf_{ij} \times \log\left(\frac{N}{df_i}\right)$$

where, $\mathrm{W}\left(t,\vec{d}\right)$ is the weight of feature t in text $\vec{d}$, N is the total number of texts, tf_{ij} is the number of occurrences lexical j in document i that is included in text $\vec{d}$ (word frequency), idf_i is the result of the total number of files divides the quotient of the number of files containing the word, and then uses the quotient to get the logarithm.

However, in our model, which aims to pick up the uncommon words, we must take these words that almost appear once into account. It is suitable to add the parameter in IDF to get the ideal result. Instead of the traditional TF × IDF, the following formula helps to resolve the effect of IDF zero value, which calculates TF and IDF according to the parameter settings. According to the SMART concept in IR, in our paper, the $\mathrm{W}\left(t,\vec{d}\right)$ can be given as follows:

$$W\left(t,\vec{d}\right) = \frac{tf\left(t,\vec{d}\right) \times log\left(\frac{N}{n_t} + \sigma\right)}{\sqrt{\sum_{t\in\vec{d}}\left[tf\left(t,\vec{d}\right) \times log\left(\frac{N}{n_t} + \sigma\right)\right]^2}}$$

where, $W\left(t,\vec{d}\right)$ is the weight of the feature t in the text $\vec{d}$, $tf\left(t,\vec{d}\right)$ is the number of times the feature t appears in the text (word frequency), and N is the total number of texts. n_t is the number of texts containing the feature t in N, $\sigma = 0.01$.

After calculating, we can find that these words, which has a low frequency but is vital to distinguish whether it contains false propaganda are still in the list of feature words.

2.2.2 Feature selection

In the model, on the one hand, the feature weight matrix in each training text is selected as the text representation type for the next text classification; on the other hand, when extracting the keyword list, according to the feature weight calculated in the previous step, the weight is selected. A feature item greater than zero, and counts the number of occurrences of feature items in the text belonging to the false promotion category in the training text, and then sets a threshold value (including the number of occurrences of the feature item and the length of the character) to count characteristic words larger than the threshold value, and is screened by a computer. Instead of artificially defining a list of false propaganda sensitive words, and comparing the list of sensitive words in the relevant national regulations to determine false propaganda, the comparison judges that the sensitive words obtained by computer processing are consistent with the false propaganda sensitive words stipulated by laws and regulations, and can be used as text search. Sensitive words determine the keywords of false propaganda.

In our model, the last and the most crucial component is to identify if the text is falsely publicized or not. If the text of web that the user browses has false propaganda, our model filters the judgment result. For the classification of texts, the traditional research mainly adopts a classification method of machine learning for category judgment. Such classification method has limitations, and the judgment of false propaganda needs to be considered rather than missed judgment. Our model is proposed to integrate multiple classification methods and vote to determine whether there is false propaganda in the text, which greatly reduces the risk of missed judgment and improves the accuracy of classification. The implementation will be detailed in next section.

3. TEXT FILTERING APPROACH

In our model, there are five kinds of sensitive word filtering methods according to text filtering based information filtering method, text clustering based information filtering method and keyword search based information filtering method. On the result of whether the text to be detected finally determined by the system is false publicity, a voting mechanism that combines five filtering methods is adopted, and the category with the largest number of votes is selected as the final category result of the text.

In our paper, we used five different classifiers, which includes supervised learning methods (Naive Bayesian, Support Vector Machines, k-Nearest Neighbor algorithm), unsupervised learning method (K-means algorithm) and Sensitive keyword search algorithm. Then, based on the result of classifiers, we chose voting strategy to select the final filter result.

3.1 Text Classifier

3.1.1 Supervised learning algorithm

In our model, it is necessary to judge whether the unknown text is false propaganda. The supervised learning classification algorithm can mark the category of the training sample data and cross-learn the test to make the sample quantity more sufficient, so as to more accurately determine the text category.

The naive Bayes

The naive Bayesian classification method [10] adopted by the system is based on a polynomial model, which calculates the conditional probability of the text to be classified, and finally compare The text belongs to the conditional probability size of the two categories, and the larger the probability is the text category to which the text belongs under the naive Bayes classifier.

In our paper, according to the test result, the posterior probability threshold of the category determination is adjusted to 25%, that is, the probability that the text belongs to a certain category exceeds 25%, and the text can be determined to belong to the category.

kNN

This method calculates the distance between the text to be classified and each known text, takes the k neighbors with the shortest distance, and judges the classification result of the item to be classified according to these “neighbors” [11]. In our paper, in order to balance the neighbor distribution, we used the SWF rule decision, which calculates the sum of the nearest k neighbors and the similarity of the text to be classified, reducing the adverse effects on the classifier caused by the asymmetry of the neighbor classification (uneven sample distribution).

SVM

As a two-classification method, it constructs the largest linear classifier on the feature space [12]. During establishment process, the data interval on the feature space can be maximized, which transforms into a quadratic solution problem in the planning domain. In our paper, we found a linear classifier to divide the data set into two categories, using the data point and representation category.

3.1.2 Unsupervised learning algorithm

The method of supervised learning is to identify things, and the result of the recognition is to label the data to be identified. In fact, there are some test texts which do not have clear category identifier, we used to choose the unsupervised learning to recognize by machine learning.

K-means

In the K-means clustering algorithm [13], the selection of the K initial cluster centroids has a greater impact on the clustering results, because the selection of the centroid at the beginning is random and arbitrary, and the result is generally not reasonable enough, just casually The data sets are randomly divided in order, so multiple iterations are required in order to achieve the final desired classification result. The initial centroid is regained every iteration until the final clustering result reaches the desired result.

3.1.3 Sensitive keyword search algorithm

In our model, we used the keyword list obtained by the feature selection in the previous step as the sensitive word vocabulary and to search for the sensitive words of all the word segmentation texts. The function of sensitive word lookup can be achieved by state transfer on the DFA tree structure. The basic method of implementing this DFA is the relationship between the index of the array and the value of the array. In this way, a tree structure representation of the DFA can be constructed. Each byte in the input text is traversed again, and then a state transition transformation is performed in the DFA to determine whether a certain prohibited word appears in the input text. Finally, a sensitive word search filtering function based on the determined finite state machine is implemented.

3.2 Voting Strategy

In previous studies, the determination of text classification results was basically based on one or improved classification method to optimize the judgment results. For our model, in order to avoid missing false propaganda texts, we use a method that combines multiple classifier results to formulate a voting strategy and strive to make the judgment result more accurate.

In our paper, we designed three voting strategies for the determination of text filtering results. Suppose there are k independent classifiers $C_k(k = 1,2,3,4,k)$ acting on the text x respectively, corresponding to the input x, each classifier has a label input, here we mark the false with 0 Propaganda, marked with 1 non-false propaganda, the final output of the text is marked $\mathrm{R}(x)$, with a corresponding voting strategy, which are shown as following:

$$C_k(x) = \begin{cases} 0, & fake, \\ 1, & others \end{cases}$$

(1) Veto power, if there is a classifier that believes that the text has false propaganda, we will determine that the text is false propaganda.

$$\mathrm{R}(X) = \begin{cases} 0, & \exists C_k(x) = 0, \\ 1, & \forall C_k(x) = 1 \end{cases}$$

(2) Vote by majority, if there are more than three votes for the false propaganda, the text is considered to have false propaganda, $count(k)$ is number of classifiers.

$$\mathrm{R}(X)=\begin{cases}0, & \exists C_k(x)=0\cup count(k)>2,\\ 1, & \exists C_k(x)=1\cup count(k)>2\end{cases}$$

(3) Full vote, only if all classifiers believe that there is false propaganda, can we judge that the text has false propaganda.

$$\mathrm{R}(X)=\begin{cases}0, & \forall C_k(x)=0,\\ 1, & \exists C_k(x)=1\end{cases}$$

4. EXPERIMENTAL EVALUATION

4.1 Dataset and Experimental Setting

In the acquisition of the text of the webpage, the webpage text content of the website is obtained by crawling according to the fake publicity food and regular food website regularly published on the website of the State Food and Drug Administration. In our experiment, we used the existing mature "Octopus" data collector to directly crawl the text on the webpage, while the latter needs to use the image recognition tool to identify the text in the image, thereby obtaining the final Web page text. There are 180 web pages, including 120 webpage texts with false advertisements and 60 legal web page texts (no fake advertisements).

In our model, 70% of the webpage text is randomly selected as the training sample, and 30% of the webpage text is used as the test text. In order to avoid the superiority of classifier to training set only, we adopt cross-validation without specifying training set and test set.

For the evaluation index, the main parameter indicators include Accurate, Precision, Recall and F1. The evaluation function design here is embodied in how to calculate these four indicators. Our model used the above-mentioned methods in the evaluation of naive Bayes classification, kNN classification, SVM classification text classifier, and text clustering and text keyword search. Four indicators are used to evaluate the text filtering effect.

4.2 Evaluation and Results Discussion

For the three voting strategies mentioned above, the experiment was evaluated by a combination of the correct rate of the text classification and the error rate. For our experimental model, different voting strategies will produce different categories of judgment results. When the voting strategy makes the false publicity judgment result high and the error rate is low, we can select the best voting method. For the calculation of the correct rate, we can find the previous section.

4.2.1 Text classification results

In our experiment, the naive Bayesian classification algorithm was used to test 54 texts and 38 texts correctly classified. After the threshold is adjusted, the naive Bayesian classification algorithm tests 54 texts and correctly classifies 43 texts, 54 texts were tested using the K-proximity algorithm, and 35 texts were correctly classified. Using a support vector machine algorithm to test 54 texts, 28 texts correctly classified. The clustering algorithm was used to test 54 texts and 43 texts correctly.

From Table 1, for the comparison of the evaluation results of four automatic classifications of text information filtering modules, the following three conclusions can be obtained:

1) For accurate, the Naïve Bayes algorithm and the K-means text clustering algorithm have the highest accurate, and the Nass Bayes correctness rate after adjusting the threshold is obviously improved, while the K-proximity algorithm and the support vector machine algorithm The accurate and precision are relatively low compared with the other two automatic classification algorithms. The vector machine algorithm finds the matching point of distinguishing the two categories of text according to the characteristics of the training sample data itself, which is difficult, and the probability that the classified data is disturbed is relatively large.

Table 1. Results of different classifiers

Classifier	Native Bayes	KNN	SVM	K-means
accurate	0.71	0.65	0.62	0.81
precision	0.70	0.69	0.69	1.00
recall	1.00	0.78	0.78	0.73
FP rate	0.92	0.72	0.72	0
TP rate	0	0.22	0.22	0.27
F1	0.82	0.73	0.73	0.82

2) For precision and recall, the four automatic text classifiers in the above figure are both high, and the purpose of constructing an optimal classifier is to construct a classifier with an ideal accuracy and recall rate. The classification results shown in the above figure are ideal for solving the problem of whether the judgment text has false propaganda.

3) For the false positive rate and the true positive rate, among the four text classification algorithms shown in the above figure, except for the cluster text algorithm, the false positive rate of other automatic classification algorithms is obviously positive and positive, and the judgment of false propaganda Therefore, even if there is a mistake, it should be better to misjudge the legal text and not let it go illegal. Text, so the above classification algorithm has a better judgment effect.

4.2.2 Voting results

In our model, we also used a keyword search algorithm based on sensitive word list, and the difference and automatic classifications technology, using DFA to construct the search tree to achieve the filtering algorithm. The rate is higher, up to 96%.

On the whole, we can see from the Figure 2, the accuracy of the judgment results by voting is very high, and the electoral system of the five algorithms is combined to ensure the correctness of the false propaganda judgment. As it is shown in Figure 2, three voting strategies have different accurate and error rate.

1) For Veto power, although it has a high accurate, the error rate is also very high. This shows that while judging false propaganda, he often uses a single classifier to determine the final result, making the probability of misjudgment high. The method obviously does not apply to our model.

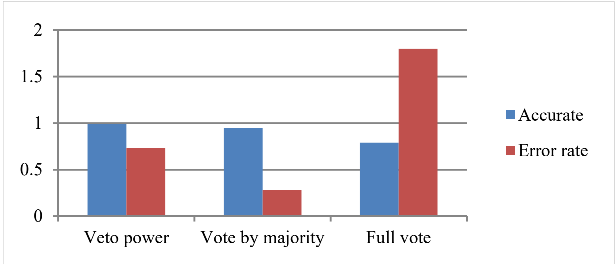

Figure 2. Results of different voting strategies

2) For Vote by majority, the principle of obeying the majority, we can see that the accurate of this strategy is very high, the error rate

is very low, indicating that when determining the final category of the text, reference to all the classifier results, not affected by a single decision, is a very good voting election strategy.

3) For Full vote, the whole vote can only be used to identify the text as false propaganda. Such voting method will reduce the accurate of judgment. For our model, we prefer to make a wrong judgment and not to miss the judgment. This method is too strict and it misses some false propaganda text strategy.

In summary, in our model, we believe that the second voting strategy has a suitable accurate and error rate, which is in line with our model needs.

5. CONCLUSION

In this paper, we proposed a sensitive word filtering model, which aims to identify webpage text with false promotions automatically. First, we used given tools to get these test texts from food and drug websites. Then we began to do text preprocessing by using stop words list. After that, Chinese word segmentation is used to cut these sentences into small segments. In the text processing module, the TF-IDF weight calculation method is adjusted to achieve the purpose of processing the food information targeted in the text, and the selection of the sensitive words for the false propaganda is obtained by designing the algorithm to filter the extracted features. Before classing these texts, we extracted features instead of words for calculating its TF-IDF weight. Finally, based on these above, we combined the automatic classification technology of machine learning uses the accuracy, recall rate, F1 value and model training time to judge the classification algorithm effect and classification performance, and supplements the judgment with the method of sensitive word search. The final decision of the five classification results is adopted through the voting election mechanism. The category gives the ideal filtering result and shows the effectiveness and efficiency of the proposed methods. In future work, we will try to optimize the multiples of the K-means clustering algorithm to evaluate our test model.

6. ACKNOWLEDGMENT

The work was partially supported by the National Key R&D Program of China under Grant No. 2017YFC0803300, National Natural Science Foundation of China under Grant No. 61703013, 91546111 and 91646201, Beijing Municipal Education Commission Science and Technology Program under Grant No. KM201810005023, KM201810005024, KZ201610005009), the Construction Project for National Engineering Laboratory for Industrial Big-data Application Technology (312000522303) and the Open Research Fund Program of Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data.

7. REFERENCES

[1] Michael Wessel, Ferdinand Thies, and Alexander Benlian. 2016. The emergence and effects of fake social information: Evidence from crowdfunding. *Decision Support Systems* 90 (2016),75–85

[2] Yu Suzuki and Satoshi Nakamura. 2018. Information Filtering Method for Twitter Streaming Data Using Human-in-the-Loop Machine Learning. In *International Conference on Database and Expert Systems Applications.* Springer, 167–175.

[3] Muhammad Asif Hossain Khan, Masayuki Iwai, and Kaoru Sezaki. 2013. An improved classification strategy for filtering relevant tweets using bag-of-word classifiers. *Journal of Information Processing* 21, 3 (2013), 507–516.

[4] Xing Wang, Jun Xie, Linfeng Song, Yajuan Lv, and Jianmin Yao. 2013. Phrase Filtering for Content Words in Hierarchical Phrase-Based Model. In *Workshop on Chinese Lexical Semantics.* Springer, 490–498.

[5] Amal Babour and Javed I Khan. 2014. Tweet Sentiment Analytics with Context Sensitive Tone-Word Lexicon. *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence*, Vol.1.IEEE, 392-399.

[6] Antoine Briand, Sara Zacharie, Ludovic Jean-Louis, and Marie-Jean Meurs. 2018. Identification of Sensitive Content in Data Repositories to Support Personal Information Protection. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems.* Springer, 898–910.

[7] Anne-Laure Bianne-Bernard, Fares Menasri, et al. 2011. Dynamicand contextual information in HMM modeling for handwritten wordrecognition. *IEEE transactions on pattern analysis and machine intelligence* 33, 10 (2011), 2066–2080.

[8] Kewen Chen, Zuping Zhang, Jun Long, and Hao Zhang. 2016. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications 66 (2016),* 245–260.

[9] Lu Liu and Tao Peng. 2014. Clustering-based Method for Positive and Unlabeled Text Categorization Enhanced by Improved TFIDF. *Journal of Information Science & Engineering* 30, 5 (2014), 1463–1481.

[10] Diab M Diab and Khalil M El Hindi. 2017. Using differential evolution for fine tuning naive Bayesian classifiers and its application for text classification. *Applied Soft Computing* 54 (2017), 183–199.

[11] Fatiha Barigou. 2018. Impact of Instance Selection on kNN-Based Text Categorization. *Journal of Information Processing Systems* 14, 2 (2018).

[12] Zhongwei Sun, Keyong Hu, Tong Hu, Jing Liu, and Kai Zhu. 2018. Fast MultiLabel Low-Rank Linearized SVM Classification Algorithm Based on Approximate Extreme Points. *IEEE Access* 6 (2018), 42319–42326.

[13] Xianmei Lang, Zairang Zhao, and Guixi Xiong. 2017. The Analysis of Traffic Drivers' Behavior based on Kmeans. In *Proceedings of the 2017 International Conference on Computer Science and Artificial Intelligence.* ACM, 232–236.