Hindawi
Mobile Information Systems
Volume 2021, Article ID 4206424, 13 pages
https://doi.org/10.1155/2021/4206424

Research Article

Data Mining Technology Application in False Text Information Recognition

Jie Wan,[1] **Xue Cao**,[2] **Kun Yao**,[3] **Donghui Yang**,[2] **E. Peng**,[1] **and Yong Cao**[3]

[1]*Fundamental Space Science Research Center, Harbin Institute of Technology, Harbin, Heilongjiang Province 150001, China*
[2]*School of Economics and Management, Southeast University, Nanjing, Jiangsu Province 211189, China*
[3]*Department of Mechanical Engineering & Automation, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, Guangdong Province 518055, China*

Correspondence should be addressed to Kun Yao; 19b353009@stu.hit.edu.cn and Donghui Yang; dhyang@seu.edu.cn

Received 27 September 2019; Revised 30 December 2020; Accepted 28 January 2021; Published 11 February 2021

Academic Editor: Alessandro Bazzi

False information on the Internet is being heralded as serious social harm to our society. To recognize false text information, in this paper, an effective method for mining text features is proposed in the field of false drug advertisements. Firstly, the data of false drug advertisements and real drug advertisements were collected from the official websites to build a database of false and real drug advertisements. Secondly, by performing feature extraction on the text of drug advertisements, this work built a characteristic matrix based on the effective features and assigned positive or negative labels to the feature vector of the matrix according to whether it is a fake medical advertisement or not. Thirdly, this study trained and tested several different classifiers, selected the classification model with the best performance in identifying false drug advertisements, and found the key characteristics that can determine the classification. Finally, the model with the best performance was used to predict new false drug advertisements collected from Sina Weibo. In the case of identifying false drug advertisements, the classification effect of the support vector machine (SVM) classifier established on the feature set after feature selection was the most effective. The findings of this study can provide an effective method for the government to identify and combat false advertisements. This study has a certain reference significance in demonstrating the use of text data mining technology to identify and detect information fraud behavior.

1. Introduction

In recent years, with the rapid development of the Internet and the increasing number of Internet users, false information has begun to spread rapidly and become more serious. False advertising is a typical example of this phenomenon [1, 2]. False drug advertisements, which not only damage the legitimate rights and interests of patients but also lead to the loss of property or life, have inflicted serious harm on society. The Local Administration for Market Regulation (AMR) for every Chinese province has formulated corresponding management measures such as setting up false advertising monitoring systems or illegal drug advertisement exposure columns on their official websites to regularly publish illegal and false advertising announcements. However, an increasing number of drug advertisements have been propagated through Internet. Many websites that publish fake drug information have been able to return to the market through re-registration. On the other hand, the various forms of drug advertising on the Internet are not limited to medical websites. They can be hidden in medicine-related post bars, forums, publicity microblogs, and promotion platforms. At the same time, the current legal system related to Internet information services is not sound enough in terms of false drug advertisements. It is difficult to supervise drug advertisements as it is hard to investigate false information and conduct follow-up tracking on the responsible institutions. Furthermore, some investigated false advertisements continue to be survived after some modifications.

Technically, the first step of cracking down on fake advertisements and preventing their resurgence is to identify

them effectively. With the development of machine learning and artificial intelligence algorithms, data mining technology has been effectively applied in many practical classification and regression problems [3], such as medical image recognition [4], checks on journal impact factor manipulation [5], wind speed prediction [6], and identification and prediction of energy efficiency factors [7]. For example, DeepMind AlphaZero not only plays Go but has also learned to sweep chess and Japanese chess. It highlighted the fact that a single algorithm that can solve multiple complex problems is an important step in creating a general machine learning system capable of solving practical problems. At the same time, this also helps to realize the development of various comprehensive decision support systems. In particular, credit card fraud detection is one of the most typical applications of these systems [8–10]. There are both supervised and unsupervised methods of fraud detection. Supervised methods use class label records in fraud or real samples to model and tag the category attributes of the new records. It is more effective in classifying types of fraud that have already occurred and performs less well on new types. Unsupervised methods do not provide class labels but instead look for anomalous data to cluster.

With regard to supervised fraud detection methods, McLachlan et al. have demonstrated that traditional statistical models, such as linear discriminant analysis and logical discriminant analysis, are very effective classification tools for fraud detection in many fields [11]. However, neural networks, C4.5, and other more powerful algorithms have been increasingly used in fraud detection with the progress in science and technology [12–14]. For the unsupervised fraud detection algorithm, the data samples do not need to be assigned to the category attribute. The association analysis and the similarity analysis methods have been used to establish a collection of fraud cases that reach a certain degree of similarity. The process of judging new cases is the process of classifying different sets. Another method commonly used in fraud detection is hierarchical clustering analysis.

In the field of fraud detection, in-depth and diverse research on credit card and telecommunication fraud has been conducted [15, 16]. In fact, these traditional fraud detection methods have also been applied in the research of false medical information. Adrienne performed an analysis of over-the-counter drugs and prescription drug advertisements broadcast on a TV station from 2008 to 2010. These advertisements were classified into objective real advertisements, suspected false advertisements, and false advertisements. Suspected false advertisements generally miss important information, exaggerate facts, relate to lifestyle, or express opinions. False advertisements provide de facto false or unconfirmed information. The conclusions of the study indicated that suspected false advertisements aimed at patients are very common for both prescription and over-the-counter drugs, and the social value of medical advertisements is contrary to the goal of providing patients with reliable drug information [17]. However, the effectiveness of methods based on text feature mining technology applied to recognize network false drug information has not been conclusively reported by a large number of authoritative studies in the literature.

In this study, we attempt to address two questions: (1) What kind of features to be collected to portray false drug advertisement? (2) Which machine learning methods are fit to improve the accuracy of recognizing false drug advertisement? These issues are all explained in this article.

In this paper, a text feature mining technology application in network false drug information recognition is presented. The sections of the paper are arranged as follows. In Section 2, the basic theories including the necessity of identifying false drug information, the text classification method, the feature selection method, and the algorithm performance evaluation method are introduced. Section 3 presents data sorting and feature selection, including the collection of fake drug advertisements notified by the China Food and Drug Administration (CFDA) and Chongqing AMR and the granted approval numbers of the drug advertisement data. Then, databases composed of false and real ads were created. Chinese word segmentation and labeling were applied to the text content of drug advertisements in the database. Feature selection and ranking were performed based on the information gain method, and a new feature set was generated through this selection. Section 4 introduces four typical classification algorithms. We used Weka to train and test several classifiers, evaluate them with certain indicators, and select the best performing classifiers to use as the model for the recognition of false drug advertisements. Furthermore, the key features of the classifications were determined. Section 5 uses the selected classifiers to predict the class labels of new data collected on Sina Weibo. This step can verify the scalability and validity of this classifier, which establishes the algorithm foundation for the construction of the decision support system. Section 6 presents the conclusions of this study and the prospects for future research.

2. Summary of the Basic Theory

2.1. Identification Necessity of False Drug Information. In China's Provision for Drug Advertisement Examination, a drug advertisement refers to any advertisement published through various forms of media that contains drug name indications (functions) or other relevant content. Such advertisements must be examined and approved in accordance with these provisions. The properties of pharmaceutical advertisements are different from those of general advertisements because of the professionalism they require. These ads not only disseminate medication knowledge for patients but also provide pharmaceutical companies with effective publicity. Thus, it is a great promotional method for new drugs on the market. However, with the propagation of false advertisements, weak supervision of network information, and asymmetrical pharmaceutical market information, imperfect laws and regulations have increased the possibility of patients being misled by false medical advertisements, which in turn has caused serious harm. Firstly, false advertising has the potential to worsen the patient's

condition and obstruct the optimal hospitalization time, which has the potential to increase the patient's pain. In such instances, it may be unavoidable that it is too late for some patients when they are finally sent to a regular hospital. Secondly, false drugs tend to increase the patient's consumption. These advertisements advise patients to increase their dosage according to a certain course of treatment. This can put a heavy financial burden on their family. Thirdly, false information undermines normal and stable market competition. Many standard hospitals have lost patients to beauty salons, folk medicine establishments, and other places that are ineffective in treatment. Finally, false advertisements may result in serious medical accidents, which not only endanger the patients themselves but also cause social concern by undermining hospitals' reputations and the government's credibility. Although false advertisements have ever-changing visual formats and publishing methods, their words and logic possess certain characteristics such as language features, tone, and content. Even though fake websites often change their facade, because the advertisements are selling the same medicine, the content logic is still traceable. However, the large variety of electronic texts distributed on the Internet with factors such as various types, distribution skew, complex relationships, frequent updates, and difficult labeling has posed significant challenges for text classification in recent years. Therefore, the effective identification of false medical information is an important research topic.

At present, with the development of big data technology and artificial intelligence algorithms, a false drug advertisement recognition model has been established with strong classification capabilities. However, its accuracy and efficiency in advertisement identification can be greatly improved, and cumbersome artificial cognition can be decreased as well. The work focuses on extensive identification, and it can be handed over to manual identification if the results are poor. After these two steps, the task can be completed quickly and accurately.

2.2. Text Classification. The main task of text classification is to clearly label the contents of new text according to a pregiven text dataset combined with samples and their class labels. It has been widely used in the fields of natural language processing and understanding, information organization and management, and content-based filtering. Sebastiani has summarized the development of text classification and correlation techniques [18].

Text classification consists of two processes: training and testing. The first step is to generate the training and testing data, which means reducing the feature dimensions through the feature selection algorithm and selecting the feature subset that best represents the data as a whole. The next step is to train and test the selected classifier and evaluate it based on the classification results of the test. In the process of English text classification, the primary way is to generate features through n-gram, but in Chinese, the primary method is word segmentation.

2.3. Feature Selection. The feature dimension of text classification data is typically very large. Its time-space complexity is high unless some feature selections are made. Feature selection focuses on expressing data with fewer features, but the selected feature set must ensure the classifier's performance [18–21]. Common methods include document frequency (DF), information gain (IG), mutual information (MI), χ^2 statistic, and term frequency-inverse document frequency (TF-IDF). Where c denotes a category, $w = \{w_1, w_2, \cdots, w_n\}$ is a term subset within the training set.

2.3.1. Document Frequency. Document frequency is the simplest feature selection algorithm. It determines how many texts contain a certain word in the entire dataset, and DF is calculated for each feature in the training set. Features are removed based on preset thresholds if the DF is particularly low or high.

$$\mathrm{DF}(w_i, c) = p(w_i|c). \tag{1}$$

where $c = \{c_1, c_2, \cdots, c_n\}$ denotes a categories. w_i is a term in a term subset $w = \{w_1, w_2, \cdots, w_n\}$ within the training set.

The biggest advantage of DF is its fast speed. Its time complexity is linear with the number of texts, which means that this method is suitable for the feature selection of large datasets.

2.3.2. Information Gain. Information gain is the difference in information entropy before and after the appearance of a feature in the text. The IG considers the feature information representation in the text category when the feature appears or not. The feature weight is characterized by the amount of information and then the filtered features. Typically, we choose a large IG value, which means that the features with a higher classification contribution constitute a classification feature subset to improve the efficiency of the system. IG can also be exploited to identify key features in the classification system.

The evaluation function of IG is as follows:

$$\begin{aligned}\mathrm{IG}(w_i) &= \sum_{i-1}^{m} p(c_i)\log_2\left(\frac{1}{p(c_i)}\right) \\ &= p(w)\sum_{i-1}^{m} p(c_i|w_i)\log_2\left(\frac{1}{c_i|w_i}\right) p(c_i|\overline{w}_i)\log\ 2\left(\frac{1}{c_i|\overline{w}_i}\right). \\ &= p(\overline{w})\sum_{i-1}^{m}\end{aligned} \tag{2}$$

2.3.3. Mutual Information. Mutual information is used in the field of feature selection to calculate the dependence severity between feature w and category c. Feature weight is defined as the feature w merged with the MI of each class. The more MI between w and c, the more authentication information related to c contained in t. The larger the MI, the greater the contribution of c and w.

The w and c mutual information is defined as

$$\begin{aligned} \mathrm{IG}(w_i) &= \sum_{i-1}^{m} p(c_i)\log_2\left(\frac{1}{p(c_i)}\right) \\ &= p(w)\sum_{i-1}^{m} p(c_i|w_i)\log_2\left(\frac{1}{c_i|w_i}\right) \\ &= p(\overline{w})\sum_{i-1}^{m} p(c_i|\overline{w}_i)\log\ 2\left(\frac{1}{c_i|\overline{w}_i}\right). \end{aligned} \tag{3}$$

2.4. Algorithm Performance Evaluation. Classification performance evaluation is used to select the best classifier from the classification experiment. In practice, indicators such as F-measure, Recall, and Accuracy are always used to evaluate the classification model. Different indicators have different evaluation significance for classifier performance, so it is necessary to select them specifically [22].

2.4.1. Recall and Accuracy. The higher the Recall, the less the text that the classifier misses on a given category, which indicates that the classifier has good performance. The higher the Accuracy, the more the text that the classifier has correctly assigned to a given category. Table 1 represents text classification with an existing category.

$$p = \frac{a}{a+b}, \tag{4}$$

$$r = \frac{a}{a+c}. \tag{5}$$

2.4.2. F-Measure. In general, Accuracy indicates the classifier accuracy and Recall indicates the classifier completeness. The evaluation criteria used in an experiment depend on the user's focus. These indicators are complementary, and simply raising one of them will lead to a decrease in another. Thus, most classifiers should strike some balance between them to avoid one indicator becoming too low. The F-measure combines these two indicators for performance evaluation—its mathematical formula is expressed as follows:

$$F_\beta - \text{measure} = \frac{\left(\beta^2+1\right)\times P\times R}{\beta^2\times P\times R}, \tag{6}$$

where β is the adjustment parameter, which is used to adjust the proportion of Accuracy p and Recall r in the calculation formula. In practice, take $\beta = 1$; then get

$$F_1 - \text{measure} = \frac{P\times R\times 2}{P+R}. \tag{7}$$

3. Data Sorting and Feature Selection

This study selected illegal and fake drug advertisements collated by the CFDA and Chongqing AMR in a false drug advertisement dataset. Drug advertisement samples granted an approval number and published by the CFDA were used as the positive class. Then, the advertising sources, drug names, and names of the manufacturing companies were collected to create complete databases. At the same time, the same number of legal drug advertisements was randomly selected for comparison with fake samples. Then a dataset with all this drug advertisement samples was built such as False/True Ads DataSet in Figure 1.

According to the theoretical basis of text classification, the text features were divided into three types: lexical features (F1), syntactic features (F2), and specific content features (F3). The original feature set was obtained through text segmentation and labeling of the acquired advertisement text content. The feature set was further selected in Weka using the IG method to generate a new set.

Classifiers J48, SVM, Naïve Bayes, and NN were trained using new feature data, and this classification model was evaluated by testing data. The classification effect of each model was evaluated using indicators such as F-measure, and subsequently a false drug advertisement recognition model was obtained. Simultaneously, model key features were obtained by analyzing the classification contribution degree of each feature set, word frequency statistics, and lexical matching results. Finally, the model applies the new data and determines the Accuracy, which then can verify the validity and scalability of classification. Figure 1 maps the overall technical procedure of this study.

3.1. Data. The main duties of the CFDA include drafting laws and regulations for drug supervision, formulating drug standards, establishing classification management systems, supervising, and investigating and punishing major illegal acts. Its website (http://www.sfda.gov.cn/WS01/CL0001/) has set up a special column for medicine that focuses on exposing illegal and false advertisements. Since July 2001, the CFDA has continuously issued false advertising announcements and quarterly summary information involving thousands of pharmaceutical companies and pharmaceutical websites. A public pharmacy advertisement inquiry has also been set up on its official website. There, you can check the registration number and approval number, address, common name, trademark name, advertisement validity and content, drug category, and organization name of various registered prescription and nonprescription drugs.

The harmfulness of false drug advertisements has led local AMRs (Administration for Market Regulation) to formulate corresponding management measures. It highlights fake drug information regularly through its established false advertising monitoring systems. Since 2012, the monitoring column of the Chongqing AMR has issued more than 30 advertisement monitoring warning announcements, including the product (service) name, category, agency, and explanations of the advertising content and its illegality.

This study utilized about 484 data samples from the CFDA and Chongqing AMR public information network, including 242 examples of false advertising and 242 examples of real advertising. We used this false and real information to establish a dataset—a sample of this data is shown in Figure 2.

Table 1: Classification system.

	Number of samples that belongs to this category	Number of samples that do not belong to this category
Amount belonging to this class	*a*	*b*
Amount not belonging to this category	*c*	*d*

As shown in Table 1, the Accuracy and Recall are defined as follows:

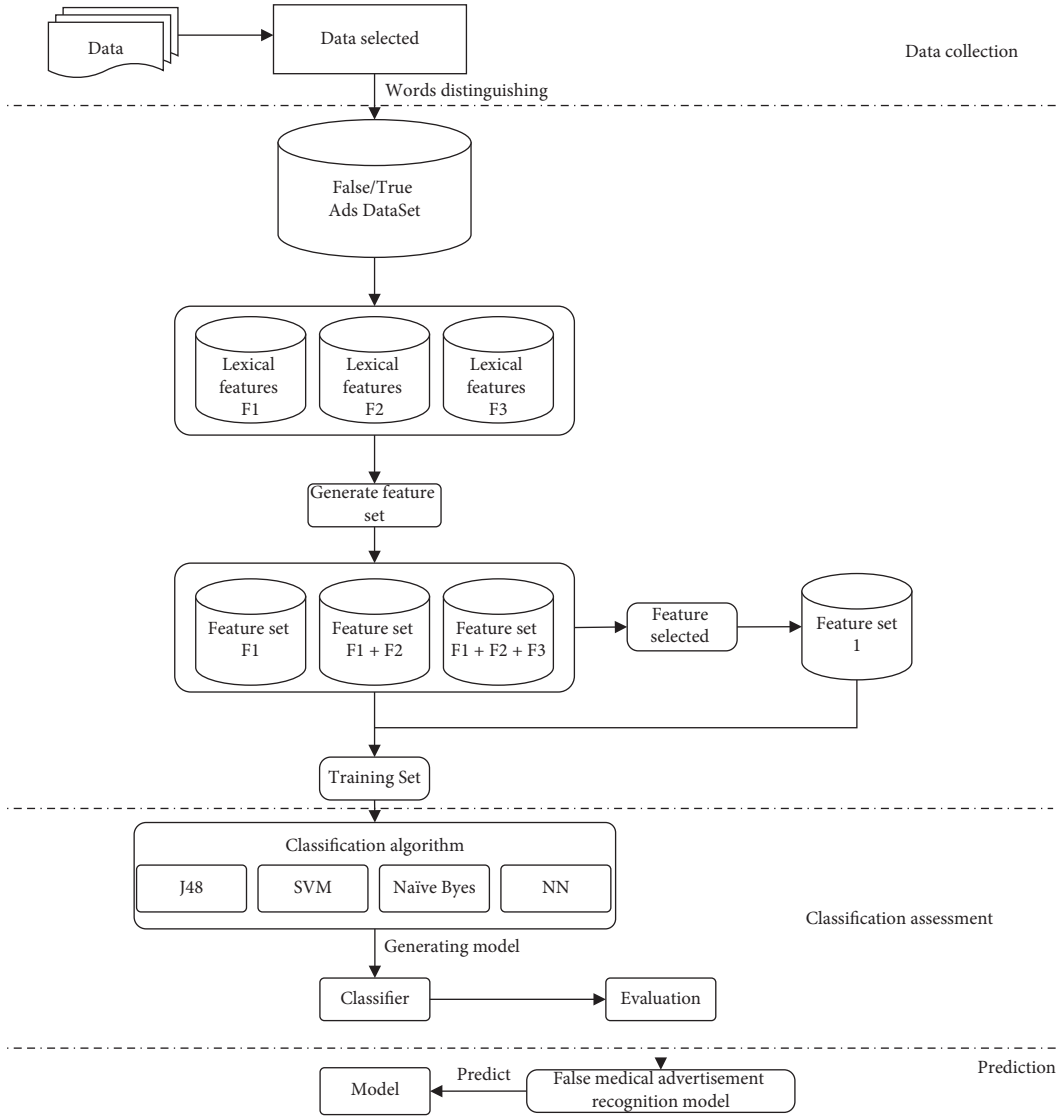

Figure 1: Technical roadmap of false drug advertisement recognition.

Although the trademark names are more familiar in the drug advertisements, the generic names can be used to indicate the exact drug name. In addition, advertisements checked by the CFDA all have an approval number. Illegal advertisements published in the column indicate the drug name, relevant reasons why it violates laws or regulations, manufacturer or publishing organization, and advertisement content, which can be used as the objects for text classification

3.2. Feature Classification. The purpose of performing content classification on drug advertisements is to find some features to best identify the data in the collected drug advertisement text. In previous research, text data features were generally divided into context-free and content-specific features. Context-free features include lexical features, syntactic features, and structural features. Structural features indicate the level and structure of an article, including whether an article contains a greeting,

False Data
Drug name: Powerful royal jelly pills
Organization name: Tianjin Central Pharm Co., Ltd.
Ads content: In 6 months, patients with heart disease did not panic, and the heart disease was completely cured. The clinical cure rate was 100%
Illegal reason: Contains unscientific descriptions of effects, expanding the cure rate or effectiveness

True data
Generic name: Compound Paracetamol and Amantadine Hydrochloride
Granted approval numbers of the drug advertisement: 2014100058
Organization name: Shandong SBOND Pharmaceutical Co., Ltd.
Ads content: In the face of challenges, I don't hesitate to fight against colds, using Ganbang, cure colds and prevent flu.

Figure 2: Sample of false/real drug advertising data.

link, or reference. Because slogans only have a few paragraphs and do not contain the basic elements of an article, the structural features are not applicable to this type. This study chose three types of features: lexical, syntactic, and content-specific.

(1) Lexical features generally include character-based features and vocabulary-based features. Besides considering English text processing such as uppercase letters, 26 alphabetic frequencies, space, and average word length, relevant features for Chinese text processing were also selected. They include nine types of features, such as the number of non-Chinese characters, number of words, number of different words, occurrence frequency of Hapax legomena and Hapax dislegomena, and average sentence length.

(2) Syntactic features, including the punctuation frequency, function frequency, and part of speech frequency, were selected for this study. These three types of features can represent the advertising text content at the syntactic level. Punctuation, such as exclamation and question marks, can generally be used to indicate more obvious attitudes. The function words that appear in this study, which are of the Chinese grammar type, have no complete meaning compared with real words. For example, consider the words "take," "be," "accord," etc. These types of words help make Chinese sentences grammatically correct and smooth. The frequency of part of speech is the frequency of the morphology, noun, and quantifier in Chinese text after part of speech labeling. In this study, 8 punctuation marks, 42 functional words, and 20 parts of speech were selected to represent syntactic features.

(3) Content-specific features relate to specific research areas and have a significant influence on false drug advertising identification. Specific keywords in the medical field include therapy, symptoms, side effects, treatments, and effects. Here, we selected the frequency of 15 specific content keywords and the total number of sentences as content-based features.

In this study, F1, F2, and F3 represented lexical features, syntactic features, and content-based features, respectively. With the help of AntConc to perform statistical analysis of the features, the features applied in this study are shown in Table 2. The words, phrases, punctuation marks, and other data given in Table 2 are all based on the Chinese format. There are certain differences between the Chinese format punctuation marks and the English ones. Here we use English format instead.

The purpose of generating a new feature set is to observe different feature types comprehensively and to use classifiers and evaluation methods for these new datasets. The accumulation strategy widely used in previous studies was utilized to generate a new feature set. Adding more complex and relevant feature combinations and continuously adding other feature sets to a given feature set made observing changes in classification indicators more useful. The first feature set generated in this study was FS1, which only included lexical feature F1. F1 and F2 were combined to create the second feature set, FS2. F1, F2, and F3 were combined to create the third feature set, FS3.

3.3. Feature Selection. To reduce the feature matrix dimension, words with a large contribution to the classifier were retained to improve the accuracy of the model and reduce running time. The smallest feature subset could be obtained by feature selection. A large amount of research conducted on text classification features has proven that IG is a better method [18]. Therefore, this study selected optimum features from the above 95 features using the IG method. The formula expressing this process is as follows:

$$\begin{aligned} \mathrm{IG}(T) = H(C) - H(C|T) = -\sum_{i=1}^{m} P(c_i)\log_2 P(C_i) \\ + P(t)\sum_{i=1}^{m} P(C_i|t)\log_2 P(C_i|t) \\ + P(\overline{t})\sum_{i=1}^{m} P(C_i|\overline{t})\log_2 P(C_i|\overline{t}). \end{aligned} \tag{8}$$

In a classification problem, C represents the category and m is the number of the category. In this study, $m = 2$, so C_1 corresponds to truthful ads while C_2 represents false ads. T is the feature and $H(C)$ is the entropy of the classification model. $H(C|T)$ is the conditional entropy of a system with a fixed feature T, which means that the label t occurs as T occurs and label $\overline{t}$ occurs when T does not. $P(C_i|t)$ indicates the probability of C_i appearing while T occurs.

With the help of Weka's feature selection function, we set IG as the selected function. The threshold was set at 0.0025, so features could be selected while $\mathrm{IG}(T) \geq 0.0025$. The selected feature sequence and feature numbers are shown in Table 3.

4. Results and Discussion

4.1. Classifier Selection. To establish a classification model for identifying false medical advertisement information, four commonly used supervised classifiers—J48, SVM, NN, and Naïve Bayes [23, 24]—were selected for training and testing in Weka.

Table 2: Feature set divided in classification experiment.

Feature			Description	Amount	Label
Lexical features	Based on content feature	(1) Amount of Chinese characters		1	F1
		(2) Total number of characters		1	
		(3) Total number of numeric characters		1	
		(4) Amount of non-Chinese characters		1	
	Based on lexical features	(5) Amount of words		1	
		(6) Different words		1	
		(7) Hapax legomena	Words that appear only ONCE	1	
		(8) Hapax dislegomena	Words that appear only TWICE	1	
		(9) Average sentence length		1	
Syntactic feature		(10–17) punctuation frequency	“, ”, “. ”, “? ”, “! ”, “: ”,”; ”, “ / ”, “ ” ”	8	F2
		(18–59) frequency of function words	put, be, about, accord, from, as, compare, include, like, make, need, possible, ‘s, get, pass, ah, yeah, oh, maybe, all, just, later, then, if, though, actually, but later, then, after, in short, until, often, feel, how, but yes, indeed	42	
		(60–79) frequency of parts of speech	n.; V.; adj.; adv.; vl; nt; pron.; nS; m; f; q; prep.; conj.; aux.v; int.; nh; W; x; vu; i	20	
Based on content feature		(80) Total number of sentences		1	F3
		(81–95) frequency of specific keywords	Treatment, symptoms, side effects, patient, function, alleviation, period, safety, health, improvement, effect, treatment, significant, recurrence, effective	15	

(1) J48 is a decision tree algorithm of C4.5 in Weka. Its core algorithm is an ID3 algorithm developed by Quinlan in 1986. Based on the divide and conquer strategy and entropy measurement, C4.5 classifies the mixed subjects according to their attributes. J48 established in Weka can be displayed in the form of a tree. Its results are easy to understand and are highly accurate. The parameters of J48 can be tuned through the functions provided by Weka.

(2) SVM is a powerful classifier that finds the decision plane with the largest class boundary in the training set. SVM can effectively solve a classification problem even with a lower data volume and can obtain a globally optimal solution. Due to its high level of performance, it is has been widely used in many studies. SVM has several different kernel functions, such as linear kernel function, polynomial kernel function, and radial basis kernel. It also has several important parameters including cost and Gamma if RBF is selected as the kernel function.

(3) In this study, NN refers to a BP neural network. Due to its special learning ability, it is very popular and could achieve better performance in many application fields. In Weka, NN was implemented with the multilayer perceptron, which provides a standard three-layer fully connected BP neural network. After setting up the GUI, the neural network can interact with the operator to get a representation of the neural network in the process.

(4) The Naïve Bayes classifier is a probabilistic classifier that ignores different feature associations through the Bayesian theory's strong independence assumption. It assumes that the appearance of each feature is completely independent of the others. It is generally used to calculate the branching probabilities of possible conditions, which is the most common method in text classification studies. Weka also provides a Naïve Bayes classifier with a default configuration that can be used directly.

4.2. Test Design and Result Analysis

4.2.1. Test Design. The comparison results of the four classifiers on four different feature sets are shown in Table 4 and Figure 3. The comparison results of the LIBSVM and SMO are shown in Table 5 and Figure 4. As a result, SVM (SMO-RBF) performs best in SVM. The classification effect of SVM and NN is significantly higher than that of J48 and Naïve Bayes. The accuracy was the highest when applying SVM on the feature set FS4 after IG, reaching 95.04%. Table 4 shows the growth rate of the F value. Next, we will discuss the results from the various aspects considered in this study such as the classifiers, feature sets, key features, and word frequency statistics.

Table 3: Feature sequence after feature selection.

Order	Feature name (feature number)
1	Number of Chinese characters (1)
2	Total number of characters (2)
3	different words (6)
4	Number of words (5)
5	Non-Chinese characters (4)
6	Auxiliary word (U) (73)
7	String punctuation (W) (76)
8	Numeral (m) (68)
9	"." (11)
10	Adverb (d) (63)
11	Noun (n) (60)
12	Average sentence length (9)
13	Verb (V) (61)
14	"," (10)
15	Temporal words (nt) (65)
16	Of (30)
17	Adjective (a) (62)
18	Total number of sentences (80)
19	Voluntary verb (vu) (78)
20	Past tense marker (33)
21	";" (15)
22	Hapax dislegomena (8)
23	Preposition (p) (71)
24	Hapax legomena(7)
25	"/" (16)
26	Patient (84)
27	Pronoun (r) (66)
28	Quantifier (q) (70)
29	Conjunction (C) (72)
30	As soon as (43)
31	All (42)
32	Total number of numeric characters (3)
33	Descriptive word (vl) (64)
34	get (31)
35	Treatment (92)
36	Effect (91)
37	Also (40)
38	Improve (90)
39	":" (14)
40	Live (34)
41	Than (24)
42	Make (27)
43	Side effect (83)
44	" " " (17)
45	features (85)
46	Safety (88)
47	From (22)
48	Interjection (e) (74)
49	Four-word phrases (i) (79)
50	Be (19)
51	Relapse (94)
52	Period (87)
53	Like (26)
54	Effective (95)
55	"!" (13)
56	Locative words (nS) (67)
57	Symptom (82)
58	Prefix (nh) (75)
59	Auxiliary (used after an adverbial) (32)
60	Then (52)
61	Health (89)
62	Significant (93)
63	rare word (x) (77)
64	Treatment (81)
65	After (44)
66	"?" (12)
67	Later (50)
68	Include (25)
69	As (23)

4.2.2. Analysis and Comparison of Classification Results

(1) Comparison between LIBSVM and SMO. As described above, SVM has different kernel functions, but a problem has no fixed relationship with it. The most widely used kernel function is RBF as it has better performance for small or large samples that have high or low dimensional data. We used the parameter optimization functions GridSearch and CVParameterSelection in Weka to optimize the four kernel functions of LIBSVM and the two kernel functions of SMO. The parameters after optimization are presented in Table 6. Figure 4 and Table 5 show the classification effect after parameter optimization.

It is evident from Figure 4 and Table 5 that the highest classification effect of the five classifier types is above 92% except for the sigmoid kernel classifier of LIBSVM. Generally, the best results are achieved on FS4, and the growth rate of the classification effect is the largest after adding the syntactic feature (F2). Among them, SMO-RBF performed best and reached the highest value in the same group of FS2, FS3, and FS4. The highest classification accuracy for FS4 was 95.0413%. In addition, the classification accuracy of LIBSVM-RBF and SMO-polyKernel was suboptimal, and there was almost no difference between the four groups. At the same time, the modeling time of SMO and that of LIBSVM are similar to each other, between 0 and 0.1 seconds

Therefore, SMO-RBF was selected as an SVM classifier to compare the other classifiers in this study.

(2) Comparison of J48, SVM, NN, and Naïve Bayes. From Figure 3 and Table 4, it is clear that NN has the best classification effect for FS1, and its F-measure value is 0.876, SVM is 0.859, J48 is 0.857, and Naïve Bayes is 0.842. For FS2, FS3, and FS4, the classification effect of SVM was the best in the group, and the best classification effect was 95.0413% on FS4. The results show that SVM (SMO-RBF) has the best effect. The classification effect of NN was the second, and its highest value (94.21%) was seen for FS3 and FS4. The effects of J48 and Naïve Bayes were the worst, with J48 at only 92.15% and Naïve Bayes at only 88.84%.

In summary, SVM (SMO-RBF) was the best classifier for this experiment, with a high classification effect value of 95.04%. Thus, the model established by SVM for FS4 was used as the classifier for identifying false drug advertisements.

Table 4: Classification results for J48, SVM, NN, and Naïve Bayes.

Classifier	Feature set	Accuracy (%)	Recall (%)	F-measure	ROC area	Growth rate (%)
J48	FS1	85.74	85.70	0.857	0.904	—
	FS2	92.15	92.10	0.921	0.935	7.47
	FS3	91.94	91.90	0.919	0.926	−0.22
	FS4	91.74	91.70	0.917	0.926	−0.22
SVM (SMO-RBF)	FS1	85.95	86.00	0.859	0.86	—
	FS2	93.60	93.60	0.936	0.936	8.96
	FS3	94.63	94.60	0.946	0.946	1.07
	FS4	95.04	95.00	0.95	0.95	0.42
NN	FS1	87.60	87.60	0.876	0.952	—
	FS2	92.36	92.40	0.924	0.957	5.48
	FS3	94.21	94.20	0.942	0.979	1.95
	FS4	94.21	94.20	0.942	0.969	0
Bayes	FS1	84.30	84.30	0.842	0.935	—
	FS2	86.78	86.80	0.867	0.953	2.97
	FS3	88.84	88.80	0.888	0.957	2.42
	FS4	88.84	88.80	0.888	0.941	0

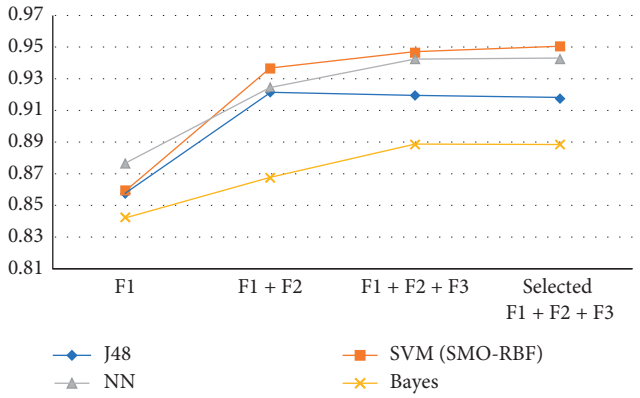

Figure 3: F-measure of classifiers in different feature sets.

4.2.3. Key Feature Analysis. Key feature extraction can help us find the most important features of fake drug advertisements, determine more precise insights into the specific field of drug advertising, and more carefully classify key features to improve the accuracy of classification in future research.

To identify the key feature types of false drug advertisements, the quantitative distribution and contribution distribution of different types of features were compared. Among the three types of features used for classification shown in Figure 5, F2 was the largest with a total of 45 features in the classification, 3 times that of F3 and 5 times that of F1. In the classification, the feature quantity contribution of F2 had a significant advantage. Figure 6 shows the quantitative distribution after each type of feature was subdivided. It was found that F2 had the highest proportion of part of speech features (60–79), with 19 eigenvectors participating in the classification, followed by 18–59 functional words, and 18 eigenvectors participating in this process. For F3, the 81–95 specific keyword frequency ranked third and had 14 eigenvectors participating in it. For the F1 results, the number of features participating in the classification was small, only 9. Its quantitative distribution was also relatively small, and the actual contribution of the classification was not too pronounced.

Figure 7 is a contribution distribution of various features in FS4 after feature extraction. The results show that F2 had the highest classification contribution at 60.60%, which is consistent with the analysis results when comparing the 4.2.2 feature set mentioned above. The classification effect after adding F2 was greatly increased. Although the quantity of F1 was relatively small, the actual contribution ranked second at 31.20%, indicating that the length of the text, richness, and other aspects are equally important for determining false drug advertisements. Although F3 was larger than the other sets, the actual contribution was only 8.20%, much lower than for the syntax and lexical features. It is evident that F2 and F1 have a very important role in classification and the quantitative distribution and contribution distribution of various features are not positively correlated.

Figure 8 shows the contribution distribution after each type of feature was subdivided. It was found that the 60–79 part of speech frequency was still the feature with the highest contribution at 37.86%, while the 10–17 punctuation frequency contribution was 11.72%. It exceeds the frequency of 18–69 functional words (11.02%) and 81–95 specific keywords (5.95%). At the same time, the contribution of F1, total number of characters, number of different words, number of words, and number of non-Chinese characters do not differ greatly as the range was between 4.1% and 4.9%. These steps then enabled us to determine the effectiveness of fake drug advertisement identification. The key feature types are part of speech frequency, punctuation, frequency of function words, and frequency of specific keywords. The frequency of a given part of speech indicates the difference in word richness. The frequency of punctuation and function words indicates the difference in the lengths of advertisement content. The frequency of specific keywords indicates the difference in the frequency of the use of specific words.

Table 5: Comparison of classification effects between LIBSVM and SMO.

Classifier		Feature set	Time (s)	Accuracy (%)	Recall (%)	*F*-measure	ROC area	Growth rate (%)
LIB	Linear	FS1	0.01	85.3306	0.853	0.853	0.853	—
		FS2	0.02	91.9421	0.919	0.919	0.919	7.74
		FS3	0.02	93.1818	0.932	0.932	0.932	1.41
		FS4	0.03	93.8017	0.938	0.938	0.938	0.64
	Polynomial	FS1	0.01	86.3636	0.864	0.863	0.864	—
		FS2	0.02	92.1488	0.921	0.921	0.921	6.72
		FS3	0.03	93.1818	0.932	0.932	0.932	1.19
		FS4	0.02	92.9752	0.93	0.93	0.93	−0.21
SVM	RBF	FS1	0.02	86.7769	0.868	0.868	0.868	—
		FS2	0.04	92.5620	0.926	0.926	0.926	6.68
		FS3	0.04	94.0083	0.94	0.94	0.94	1.51
		FS4	0.04	94.2149	0.942	0.942	0.942	0.21
	Sigmoid	FS1	0.02	82.2314	0.822	0.819	0.822	—
		FS2	0.07	79.3388	0.793	0.785	0.793	−4.15
		FS3	0.06	78.7190	0.787	0.778	0.787	−0.89
		FS4	0.06	80.3719	0.804	0.797	0.804	2.44
SMO	RBF	FS1	0.05	85.9504	0.86	0.859	0.86	—
		FS2	0.06	93.5950	0.936	0.936	0.936	8.96
		FS3	0.06	94.6281	0.946	0.946	0.946	1.07
		FS4	**0.05**	**95.0413**	**0.95**	**0.95**	**0.95**	**0.42**
	polyKernel	FS1	0	86.5702	0.866	0.865	0.866	—
		FS2	0.02	92.9752	0.93	0.93	0.93	7.51
		FS3	0.01	93.8017	0.938	0.938	0.938	0.86
		FS4	0.02	94.2149	0.942	0.942	0.942	0.43

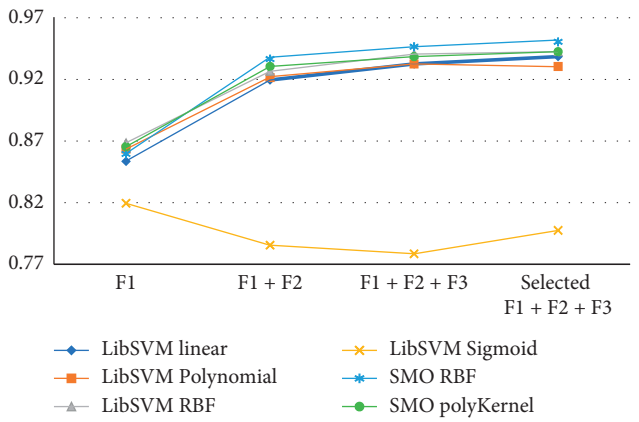

Figure 4: Comparison of classification effects between LIBSVM and SMO on different kernel functions.

5. False Medical Information Prediction

The SVM with the best classification results, as obtained from the process described in the above sections, was selected as the final classification model. To verify its validity and ability to accurately predict the category attributes of new data, it was necessary to evaluate its prediction effect. This was performed to ensure that other false drug advertisements on the network also have certain authentication functions and are expandable.

Due to the relatively fast update speed of Weibo, it has a timeliness that aligns with the characteristics of instant interactive advertising. At the same time, attention and forwarding functions are not restricted, meaning that bloggers will not only form a fixed audience but also have a wider advertising scope. As a result, Weibo is increasingly becoming the main platform for the spread of fake drug advertising. Therefore, this article uses the Sina Weibo platform as a collection platform for new data. In the Weibo search bar, the aforementioned key distinguishing signs of false advertising, such as key "completely," "take," and "one month," were entered in the "medicine" category. This yielded 33 pieces of published results from various bloggers. False advertisements for different types of drugs were mixed with 16 real drug advertisements. A total of 49 data samples were used for category prediction using the established SVM classification model. The forecast results are shown in Tables 7–9:

As shown in the forecasted total results in Table 7, the accuracy of the established classification model was 97.9592%. Table 7 shows that the accuracy of determining the false advertisements was 94.1%, the accuracy of determining the real advertisements was 100%, and the absolute average error is 0.0204. The F-measure was 0.980, and only one of the 33 false data samples available in the confusion matrix of Table 9 was judged wrong—all of the 16 true data samples were judged correctly. The results show that this classification model has high adaptability and effectiveness for use with new data and can be used not only to identify false advertisements on traditional media channels but also to efficiently identify false advertisements on the Internet.

Finally, based on the conclusions of this paper, our relevant recommendations for the supervision and governance of false drug advertisements are as follows:

(1) Strengthen the effectiveness and timeliness of drug advertisement monitoring, and improve the laws and

Table 6: Parameter optimization results of LIBSVM and SMO.

LIB	Linear	-S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -Z
	Polynomial	-G 1.0 -S 0 -K 1 -D 3 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -Z
SVM	RBF	-S 0 -K 2 -D 3 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 2.0 -E 0.001 -P 0.1 -Z
	Sigmoid	-S 0 -K 3 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -Z
SMO	RBF	-C 7.0 -L 0.001 -P 1.0E-12 -N 1 -W 1 -K -G 1.0 -C 250007 -R 1.0E-8 -M -1
	polyKernel	-L 0.3-C 2 -P 1.0E-12 -N 0 -V -1 -W 1 -K -E 1.0 -C 250007 -R 1.0E-8 -M -1

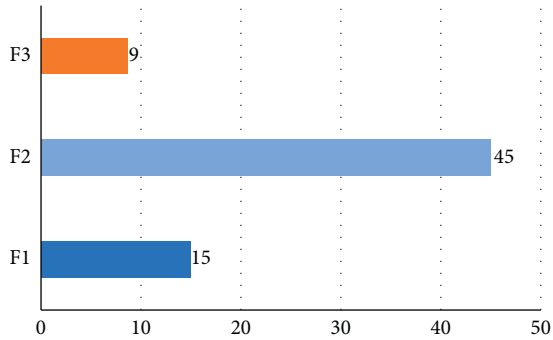

Figure 5: Quantitative distribution of F1, F2, and F3.

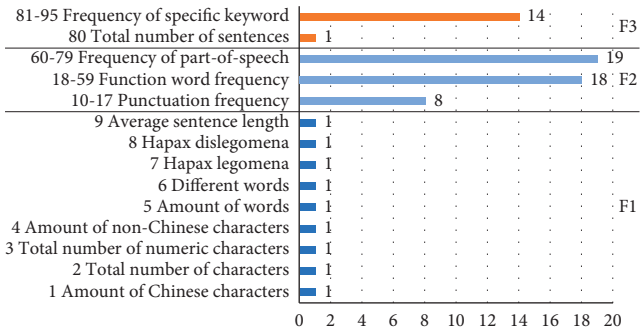

Figure 6: Contribution distribution of various features to the classification.

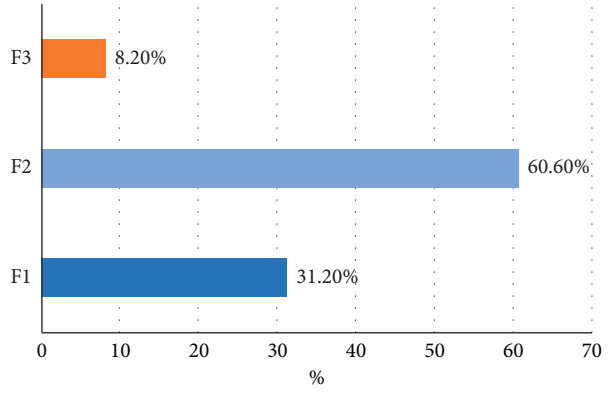

Figure 7: Classification contribution percentage of F1, F2, and F3.

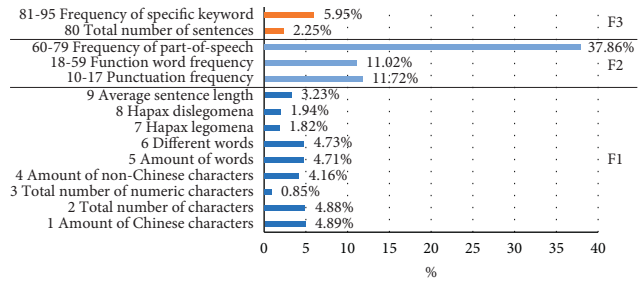

Figure 8: Percentage distribution of various characteristics for classification.

regulations on Internet information management. Using the classification model and related intelligent monitoring software established in this study, the real-time monitoring and effective identification of Internet drug advertisements on search engines, Twitter, blogs, post bars, forums, and medical websites can improve recognition accuracy and identification efficiency and reduce the tedious task of manual identification. If there are samples with unsatisfactory results, they can be handed over to manual identification, and the two processes can identify a wider range of advertisements more quickly and accurately. False drug advertisements that have been identified as exaggerated or illegal and the institutions propagating them should be publicized and severely cracked down on. At the same time, it is necessary to pay attention to tracking supervision, to prevent false advertisements from being changed and returned to the Internet in other forms. This should be done to ensure the purity of the network environment.

(2) Improve relevant regulations such as the "Public Drug Advertisement Review and Release Standard." In this study, the word frequency statistics for both the fake and real drug advertisements were collected for the entire text after word segmentation. A more in-depth analysis of commonly used words and application scenarios would be useful to provide more specific data support and a theoretical basis for drug advertising regulators or advertising law developers. This would help to improve relevant laws and regulations and to enable the relevant personnel to be more detailed and specific in devising effective laws and regulations. At the same time, this method would make it easier for administrative law enforcement personnel to operate and implement the relevant provisions or penalties.

(3) Strengthen the publicity and popularization of medicines. The conclusions in this study indicate that false advertisements generally have more obvious characteristics. With knowledge of some key features, ordinary people can identify many fake drug

Table 7: Total forecast result.

Correctly classified instances	48	97.96 (%)
Incorrectly classified instances	1	2.04 (%)
Total number of instances	49	
Kappa statistic	0.9543	
Mean absolute error	0.0204	
Root mean squared error	0.1429	

Table 8: Detailed accuracy by class.

Class	TP rate	FP rate	Precision	Recall	F-measure	MCC	ROC area	PRC area
TRUE	1.000	0.030	0.941	1.000	0.970	0.955	0.985	0.941
FALSE	0.970	0.000	1.000	0.970	0.985	0.955	0.985	0.990
Weighted avg.	0.980	0.010	0.981	0.980	0.980	0.955	0.985	0.974

Table 9: Confusion matrix.

	a	*b*
a = TRUE	16	0
b = FALSE	1	32

advertisements. Therefore, when implementing their own supervisory functions, regulatory authorities can conduct legal publicity and educational campaigns targeted at the public to provide public identification methods for false advertisements, basic medical knowledge, and rights protection knowledge. This can increase public awareness and recognition of and immunity to fake drug advertisements. Furthermore, the personnel and organizations that provide false publicity materials should be reported.

6. Conclusion and Prospects

Due to its serious harm, the supervision of and fight against false text information—especially, false drug advertisements—have always been the focus of government departments at all levels. However, the Internet has brought more diverse forms of medicinal advertising to people and caused a sharp increase in the difficulty of work for the relevant regulatory bodies. Therefore, the effective identification of false text information is of great significance.

This study took medical text information that was easy to obtain and used a clear class label for the research roles. Based on text data mining technology, an intelligent recognition model that can effectively classify and predict false text information was established to identify false information appearing on the Internet and improve recognition accuracy and efficiency. At the same time, with regard to drug advertisements, finding more important features, words, and application scenarios could make data support and theoretical basis for relevant laws and regulations sounder. Overall, the importance and innovation of the work conducted in this study were mainly reflected in the following aspects:

(1) This study used real and effective research data from the CFDA's official illegal false drug advertising information bulletin to construct the text database.

(2) It examined false advertising information to identify issues related to feature analysis, feature extraction of lexical and syntactic features, and three types of content and specific characteristics. It also used the information gain algorithm for sorting feature selections.

(3) After establishing four typical classification models and analyzing and evaluating the test results, the conclusion is that the classification model of SVM (SMO-RBF) on feature set FS4 has the best classification effect. Therefore, the classification and prediction model for the final false advertisement recognition was determined. At the same time, to verify that the classification model has good adaptability with new data, the prediction model of the classification model was carried out by obtaining the text data of the advertisement information on Sina Weibo. The results showed that the prediction accuracy of the classification model was also high. The conclusion is that, based on the established classification model, false advertisements published on both traditional media channels and social networks can be recognized effectively. The established recognition model can be used for a wide range of use with a certain level of effectiveness and scalability. This method greatly improves recognition accuracy and efficiency and reduces the tedious work of manual recognition.

(4) Through the analysis and extraction of the key features of false advertising information, as well as the contribution of individual features, and the word frequency statistics of the established false and real drug advertising database, we were able to better identify fake drug advertisements. Important features allow for more precise insight into specific areas of pharmaceutical advertising. Through a deeper understanding of the words and application scenarios commonly used in false and real advertising content, relevant laws and regulations can be improved. The results of the frequency of speech, punctuation, frequency of function words, and frequency of specific keywords are key features in the medical field. It is worth noting that fake pharmacies

generally have longer text lengths and higher content richness and often have a more positive description of the cure rate and efficiency of a given drug. However, there will be terms that are explicitly prohibited in the "Administrative Measures on Pharmaceutical Advertisements," and the use of numerals is higher. Real drug advertisements are generally short and succinct, with refined words and no assertions or guarantees, such as "completeness," indicating efficacy or safety.

Finally, according to the conclusions of this study, relevant suggestions can be made for the supervision and governance of false text information such as drug advertisements.

In addition, considering the complementary advantages of various data mining algorithms, the method of multi-classifier model fusion should be further studied to potentially improve the effectiveness of the comprehensive decision support model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the program of Shenzhen Technology Projects (nos. JCYJ20160226201347750 and ZDSYS201707280904031) and the National Natural Science Foundation of China (Grant no. 51577043).

References

[1] T. C. Morrison, "The false advertising of specialty medical products under the lanham act," *Food Drug and Cosmetic Law Journal*, vol. 44, pp. 265–271, 1989.

[2] L. Zhang, J. Zhou, Z. Zhao, and Z. Jing, "Treatment false advertisement in China: a tragedy," *The Lancet*, vol. 387, no. 10037, pp. 2505-2506, 2016.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[4] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. 9, pp. 60–88, 2017.

[5] D.-H. Yang, X. Li, X. Sun, and J. Wan, "Detecting impact factor manipulation with data mining techniques," *Scientometrics*, vol. 109, no. 3, pp. 1989–2005, 2016.

[6] J. Wan, J. Liu, G. Ren et al., "Day-ahead prediction of wind speed with deep feature learning," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 30, no. 5, p. 20, 2016.

[7] F. Meng, B. Li, D. Yang et al., "Energy efficiency evaluation method based on multi-model fusion strategy," *Cluster Computing*, vol. 19, no. 4, pp. 1937–1949, 2016.

[8] M. Riani, A. Corbellini, and A. C. Atkinson, "The use of prior information in very robust regression for fraud detection," *International Statistical Review*, vol. 86, no. 2, pp. 205–218, 2018.

[9] W. N. Robinson and A. Aria, "Sequential fraud detection for prepaid cards using hidden Markov model divergence," *Expert Systems with Applications*, vol. 91, pp. 235–251, 2018.

[10] S. Akila and U. Srinivasulu Reddy, "Cost-sensitive risk induced bayesian inference bagging (RIBIB) for credit card fraud detection," *Journal of Computational Science*, vol. 27, pp. 247–254, 2018.

[11] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, Hoboken, NJ, USA, 2004.

[12] H. He, J. Wang, W. Graco et al., "Application of neural networks to detection of medical fraud," *Expert Systems with Applications*, vol. 13, no. 4, pp. 329–336, 1996.

[13] M. Mulholland, D. B. Hibbert, P. R. Haddad, and C. Sammut, "Application of the C4.5 classifier to building an expert system for ion chromatography," *Chemometrics and Intelligent Laboratory Systems*, vol. 27, no. 1, pp. 95–104, 1995.

[14] R. S. Chhikara and J. Mckeon, "Linear discriminant analysis with misallocation in training samples," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 899–906, 1984.

[15] R. Wheeler and S. Aitken, "Multiple algorithms for fraud detection," *Knowledge-Based Systems*, vol. 13, no. 2-3, pp. 93–99, 2000.

[16] S. Rosset, U. Murad, E. Neumann et al., "Discovery of fraud rules for telecommunications—challenges and solutions," in *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 409–413, ACM, San Diego, CA, USA, August 1999.

[17] A. E. Faerber and D. H. Kreling, "Content analysis of false and misleading claims in television advertising for prescription and nonprescription drugs," *Journal of General Internal Medicine*, vol. 29, no. 1, pp. 110–118, 2014.

[18] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[19] H. T. Ng, W. B. Goh, and K. L. Low, "Feature selection, perceptron learning, and a usability case study for text categorization," *ACM SIGIR Forum*, vol. 31, pp. 67–73, 1997.

[20] Z. Zheng, X. Wu, R. Srihari et al., "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 80–89, 2004.

[21] Y. Xu and L. Chen, "Term-frequency based feature selection methods for text categorization," in *Proceedings of the Fourth International Conference on Genetic & Evolutionary Computing*, pp. 280–283, IEEE, Shenzhen, China, December 2010.

[22] D. H. Yang and H. X. He, "An automatic recognition method of journal impact factor manipulation," *Journal of Information Science*, vol. 37, no. 3, pp. 235–324, 2015.

[23] Q. Hu, S. Zhang, Z. Xie et al., "Noise model based ν-support vector regression with its application to short-term wind speed forecasting," *Neural Networks*, vol. 57, pp. 1–11, 2014.

[24] I. Kononenko and M. Kukar, *Machine Learning and Data Mining*, Horwood Publishing, Cambridge, UK, 2007.