Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2017 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Spring 7-19-2017

Research on Automatic Identification of Rumors in Stock Forum Based on Machine Learning

Hua Zhang
Southwestern University of Finance and Economics, 19372157@qq.com

Jun Wang
Southwestern University of Finance and Economics, juujuu0058@163.com

Yan Chen
Southwestern University of Finance and Economics, 504473715@qq.com

Jinghua Tan
Southwestern University of Finance and Economics, 595915575@qq.com

Qing Li
Southwestern University of Finance and Economics, 41761780@qq.com

Recommended Citation

Zhang, Hua; Wang, Jun; Chen, Yan; Tan, Jinghua; and Li, Qing, "Research on Automatic Identification of Rumors in Stock Forum Based on Machine Learning" (2017). *PACIS 2017 Proceedings*. 46.
http://aisel.aisnet.org/pacis2017/46

Research on Automatic Identification of Rumors in Stock Forum Based on Machine Learning

Completed Research Paper

Hua Zhang
School of Economics Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China
19372157@qq.com

Jun Wang
School of Economics Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China
juujuu0058@163.com

Yan Chen
School of Economics Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China
504473715@qq.com

Jinghua Tan
School of Economics Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China
595915575@qq.com

Qing Li[*]
School of Economics Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China
41761780@qq.com

Abstract

When rumors prevail in securities market, it is very difficult for investors to identify valid information. In the meantime, investors have much more ways to access information with the evolution of internet. But there is an overwhelming quantity of information on the Internet, the coexistence of facts and rumors, namely, "widely circulated" and "specious", yet "unconfirmed officially" vague information, makes it more difficult for investors who with limited rationality to distinguish facts from rumors. Existing studies are mainly devoted in the method of event study, namely screening rumors from "official channels" that clarified, which is neither timely efficient in terms of accessing to rumors nor providing the basis for decision-making. Traditional news has evolved into various forms of social media, including forums, blogs, micro-blogs etc., and users can not only gain quick access to more valuable and timely information, but also amplify information that embed the news effectively by participating in commenting on various social media. Dynamic information creation, sharing and coordination among Web users are exerting increasingly prominent impact on the securities market in now days. Thus, it is very necessary to study the effects of social media as online forums on the securities market. In this paper, the method of machine learning is adopted for the first time to identifying the Internet rumors automatically, and successfully in crawling massive forum data by smart computer technology. Unlike the case study and statistical sampling of rumors, this paper conduct automatic identification of Internet rumors by utilize the smart technology, thus paving the way for more in-depth analysis about the effects of Internet media on the securities market in future.

Key Words: Machine Learning, Automatic Recognition, Securities Market, Forum Rumors.

[*] Corresponding Author

Introduction

Internet rumor is one of the significant kinds of media information. The study about the media effects on the securities market originated in the field of finance. By observing the impacts of big events on the securities market, researchers began to use the method of event study to research the effect of specific news reports on securities market volatility. With the rise of Internet media and the booming of information technology, an increasing number of researchers have begun to try exploring the correlation between media information and the securities market through big data mining. This paper made a literature review and found that current research papers were more focus on demonstrating the relevance of different types of Internet information (forums, microblogs, Wikipedia, user behavior, etc.) and the volatility of the securities market. As can be seen from these studies that scholars are mainly focus on how to quantify the effects of media text information and how to build an analysis model for the correlation between Internet media information and the securities market.(as shown in Table 1)

Table 1. Representative Literature of the Effects of Different Media on the Securities Market

Category	Major Literature
Study About the Effects of Social Media on the Securities Market	Antweiler and Frank(2004), Bollen et al.(2011), T. Preis et al.(2013),H. S. Moat et al.(2013), X. Luo et al.(2013), Curme et al.(2014), Siganos et al.(2014)
Study About the Effects of News on the Securities Market	Wuthrich et al.(1998), Lavrenko et al.(2000), W.S.Chan(2003), M. A. Mittermayer and G. F. Knolmayer(2006), Tetlock et al.(2008), Wang et al.(2011), R. P. Schumaker et al.(2012)
Study of Both Above	Q. Li et al.(2014b,2015)

According to "Research Report Internet News Market in China 2016" by the China Internet Network Information Center (CNNIC), Internet users are still not conscious enough to question the authenticity of online news and it is common for them to directly repost unverified news. Statistics show that 60.3% of Internet users repost information directly without checking its authenticity, and only 25.7% of users consciously check the authenticity and accuracy before forwarding information. Since it costs nothing to forward unverified news, it is widely common that Internet users repost hot news readily as soon as they read it, which fuels the spread of false news. rumors are generated from lack of information, and also arise if information is too much, as is shown in Kapferer (2008). In the face of massive Internet information, which one should we believe? Isn't the most reliable and the most subjective ones for investors choose to believe? When making a decision, do investors tend to be other factors if they do not tend to the factors that be able to strengthen their personal views?

Unfortunately, although the Internet media are filled with so much rumors of information, studies by scholars on the rumors of the impact of the securities market is seriously lagging behind. Therefore, by taking the text information in the stock forum of eastmoney.com as basic data, and the support vector machine (SVM) as the classifier, this paper achieved in automatically identification of stock forum rumors for the first time. Better identification rate makes it convenient for investors to identify rumors, and more objective understanding of forum information, which helps avoid of confusion causes by rumors. This is the practical significance and innovation this paper aims at.

Related Work

Research Status and Trend

As one of the most important sources of risks in the financial market, rumors have not been fully studied yet as mentioned above. The paper has made a literature review on representative studies (as shown in Table 2). It suggests that sample selection for the research of rumors' impact on the securities market is still narrow, which may not give an objective and the whole picture of rumors in securities market.

Table 2. Case Study About the Effects of Rumors on Securities market

Major Literature	Sample Selection
Rose（1951）	manual collection of samples of two years' stock market rumors
Diefenback（1972）	US Securities Market Analysis Report
P. L. Davies and M. Canes（1978）	"Market Rumor", column of Wall Street Journal
J. Pound and R. Zeckhauser（1990）	stock rumors of M & A

B. M. Barber and D. Loeffler（1993）	“Inside Wall Street”, column of Business Week
I. Mathur and A. Waheed（1995）	“Inside Wall Street”, column of Business Week
P. Clarkson et al.（2006）	analysis of only 189 acquisition rumors in the Hotcopper forum.
Spiegel et al.（2010）	Israeli Internet forum rumors
Zhao, J. M., He, X., Wu F. Y., (2010)	rumor clarification announcements in stock markets of Shanghai and Shenzhen
Sui Y. P., (2015)	clarification announcements at cninfo.com.cn

The virtuality of highly advanced social media allows the public to give free voice and makes the media a “hotbed” for the breeding and spread of rumors. In the meantime, due to the super transmission efficiency of the Internet, rumors have greater effects than traditional word of mouth. Therefore, the study on the effects of rumors on the securities market cannot be just confined to rumors on traditional news media or “small sample data”. Modern social media must be combined to conduct massive data research so as to give comprehensive and timely feedback on the effects of rumors on the securities market. For this purpose, this paper attempts to realize automatic identification of forum rumors through machine learning and gather massive rumor samples, and strive to make breakthrough and innovation in the sample size of “the study of the impacts of rumors on the securities market”. This is also an inevitable trend to take “massive samples” for “the study of the impacts of media on the securities market”.

Yet, the study on modern social media will be inevitably accompanied by analysis based on big data. Take eastmoney.com for example. Just traditional methods of artificial screening or event study alone can no longer meet the needs of research, or the needs of recognizing rumors from massive text information based on logical rules and sign techniques of artificial extraction through the traditional “knowledge engineering (KE) classification system”. Rumors that obtained by artificial means, cannot screen massive data in an accurate and stable way and inevitably confront with a bottleneck of knowledge acquisition and representation, it is necessary to adopt an automated method to identify and study Internet rumors.

Free and open social networking has laid a good foundation for the spread of Internet rumors. In particular, self-publishing has made the connection between of people more virtual, stickier and more mutually influential and the content in the forums is more frequently shared and re-posted. Yet, there are just a scanty few studies on Chinese financial Internet rumors based on the “Machine learning(ML) classification system”. The rapid development of artificial intelligence makes it possible to process TB (Terabytes) above information within a short period of time and extract valuable and accurate information in time from the WEB media. Therefore, it is crucial to use the advanced capture and analysis technologies in the computer science to study the powerful technology of automatic identification and investigate the influence mechanism of rumors in social media, making it possible to sensitively and accurately capture the hidden relationship between Internet rumors and securities market volatility. This paper adopts the “Machine learning(ML) classification system” to automatically extract related classification rules from massive forum data and derive automatic text classifier. Compared with the traditional “knowledge engineering(KE) classification system”, it can greatly ease the problem of knowledge acquisition and representation.

By launching the web crawler, this paper successfully gathered text information in the stock forum of eastmoney.com, and for the first time to achieve in identifying the rumors of the stock forum automatically based on the machine learning method (SVM), which with remarkable results and important practical significance, which also providing premise and basis for the follow-up studies, and making it possible to conduct large-scale systematic analysis of the effects of Internet rumors on securities market volatility.

Design of Automatic Crawling and Identifying System for Internet Rumors

Internet rumors, which are carried in a variety of forms with different characteristics, such as forums, micro-blogs and WeChat. According to existing literature analysis, the social media to which investors pay more attention to are mainly financial forums. There are two major categories of financial forums. One is specialized financial websites, such as eastmoney.com, jrj.com.cn and hexun.com etc. that run special columns of securities information and stock forums. The other is financial sections of major portal websites, such as finance.sina.com.cn, money.163.com and business.sohu.com, which also run securities columns and forums. Thus, along the main line of “sample selection - preprocessing -

automatic identification", this section introduces the design and structuring principles of "the automatic crawling and identification system for Internet rumors".

Technical Diagram

According to the main line of technical research, "sample selection -- preprocessing -- automatic identification", this section plans the implementation road-map of research technology and forms the "technical diagram" (as shown in Figure 1).

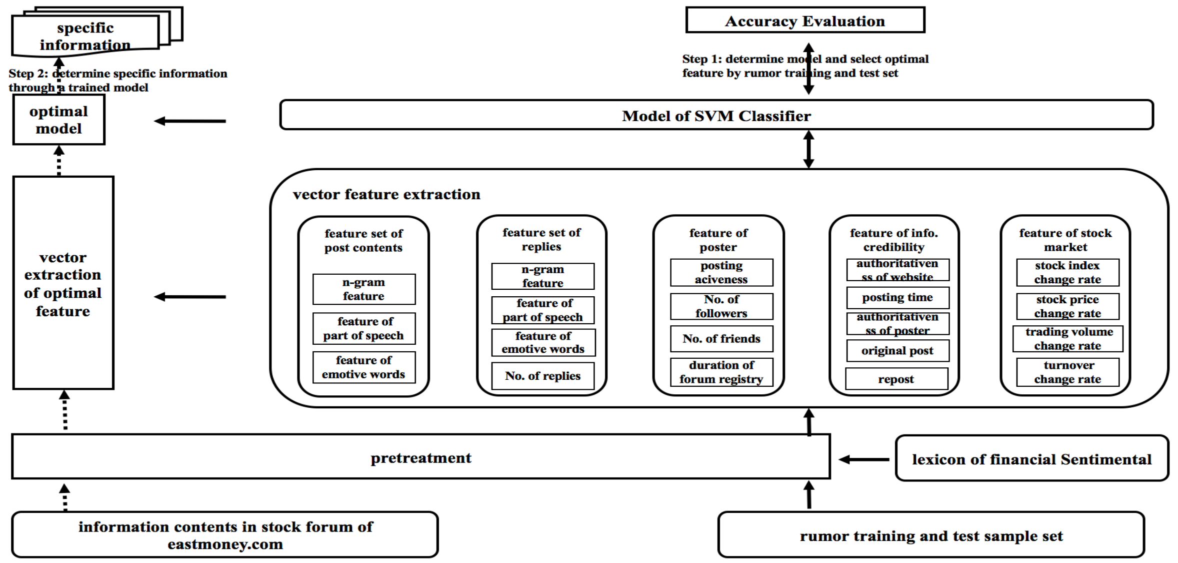

specific information
Step 2: determine specific information through a trained model
optimal model
vector extraction of optimal feature
Accuracy Evaluation
Step 1: determine model and select optimal feature by rumor training and test set
Model of SVM Classifier
vector feature extraction
feature set of post contents
n-gram feature
feature of part of speech
feature of emotive words
feature set of replies
n-gram feature
feature of part of speech
feature of emotive words
No. of replies
feature of poster
posting aciveness
No. of followers
No. of friends
duration of forum registry
feature of info. credibility
authoritativenss of website
posting time
authoritativenss of poster
original post
repost
feature of stock market
stock index change rate
stock price change rate
trading volume change rate
turnover change rate
pretreatment
lexicon of financial Sentimental
information contents in stock forum of eastmoney.com
rumor training and test sample set

Figure 1 Technical Diagram of Automatic Identification of Rumors in Eastmoney Stock Forum

Data Selection and Automatic Crawling

As an important component of the search engine's crawling system, the main purpose of the web crawler is to downloaded web pages to local copies of content mirror images, namely, to start from URLs of eastmoney.com to obtain the initial page lists and keep gathering new URLs from current pages until URLs are empty or conditions for terminating crawling are met. This paper uses the high-speed and accurate LocoySpider (an open source web crawler) to make adaptive modulation and meet the crawling requirement. Through the web crawler, the authors gathered all the posts in the stock forum of eastmoney.com, about 37 million 800 thousand posts in all, spanning from January, 2007 to December, 2016, quantity reaching 10 GBs, data precision being the second. Content gathered include: stock codes, IP addresses of posts, post titles, post contents, websites, attention rates(amount of reading), reply rates(amount of reply) and posting time.

Preprocessing of Crawled Information

First, LocoySpider is used to gather all the posts in the "stock forum" on eastmoney.com, and the gathered information is written into the MYSQL database, after which the text information is classified and derived according to the website information of each datum. Then the data of the same stock are derived into a CSV file so as to realize classification according to stock code. Second, the program is used to delete distorted information, wrong records or improper samples, such as text information with minimum (less than 4KB) or maximum (more than 100KB) capacity, or data with 0 attention rates or long suspended stocks. Through this process, about 200 thousand noise posts and 37 million 600 thousand inventory posts are removed.

Automatic Identification of Internet Rumors

When the forum text information in the "stock forum" of eastmoney.com was crawled, all the given text information is classified into predefined categories then, thus "rumors" and "non-rumors" are distinguished. The specific process of identification is “text representation - feature generation - feature extraction - text classification”.

Text representation. An Internet rumor is a piece of text information which is composed of words and punctuation, words forming phrases, sentences or passages. If the computer is to efficiently process text information, a relatively desirable method of representation is needed. As for text information representation, it’s necessary to first define the text features and then move on to the text information representation.

Step1: Text Feature Defining. The purpose of defining text features is to achieve real feedback on document content and distinguish different documents. At present, the main methods of determining text features are the bag-of-words model, the phrase model and the N-gram method. Although the bag-of-words model is relatively simple, the semantic information of text is somehow neglected. The phrase model considers the semantic information of text and is more capable of expressing the theme of the text than bag-of-words, but due to the complexity of natural language, if it fails to gain an accurate understanding of the semantic information of text, it will affect the performance of next classification. The N-gram model, a statistics-based method, can segments texts without the need of phrases segment to realize the automated processing of Chinese text information (Miller G et al., 1990). In this paper, the N-gram model is used to define text features of web text information.

Step 2: Text Representation Model. The representation of web text information is the premise and basis of automatic identification of the Internet rumors, which means to convert Internet forum texts into information that can be identified by the computer, represent it in a numerical form and describe it. Currently, the main text representation models are the Boolean Model, the Probability Model and the Vector Space Model (VSM). The Boolean Model is a precise matching model, simple and easy to operate, which is also a defect. This model is now rarely used alone. The Probability Model, based on probability and statistics, relies too much on text sets, and processes data in a too simple way. The Vector Space Model is one of the most frequently used natural language processing models (G. Salton et al., 1974), and is adopted in this paper. Its basic principles are as follows:

The paper supposes the stock forum information of eastmoney.com as *D* (Document), the feature item as *F* and the term weight as W_k, then the feature T_i exists in a certain piece of information D_i in the forum, and the feature vector of T_i is represented as 1, otherwise 0. The degree of the correlation between two documents is measured by the distance between the vectors of documents, usually calculated by the inner product or the cosine of the included angle. The smaller the angle, the greater the similarity (as shown in Figure 2).

$$Sim(D1, D2) = \sum_{k=1}^{n} w_{1k} \times w_{2k}$$

Or

$$Sim(D1, D2) = cos\theta = \frac{\sum_{k=1}^{n} w_{1k} \times w_{2k}}{\sqrt{(\sum_{k=1}^{n} {w_{1k}}^{2}) \times (\sum_{k=1}^{n} {w_{2k}}^{2})}}$$

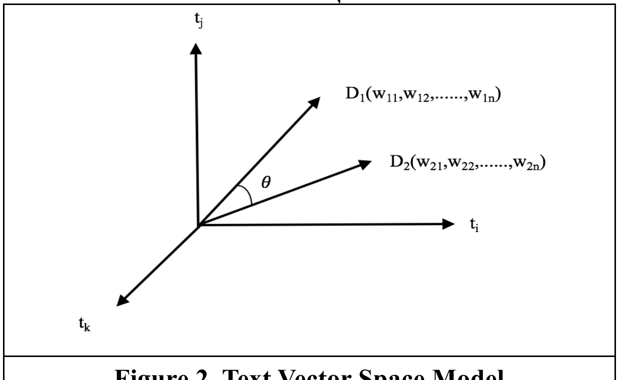


Figure 2. Text Vector Space Model

Text feature extraction. DiFonzo et al. (1994) believed that there are three key stages in the spread of rumors: generation, evaluation and dissemination. The stage of generation is mainly based on the uncertainty and anxiety of the contents. In Internet forum rumors, it is shown as the poster’s features (such as the poster’s personal characteristics, number of his or her posts and followers) and the content of the posts (such as term frequency, parts of speech and sentiment etc.). The stage of evaluation is about whether the receivers chooses to believe the rumor or not. In Internet forum rumors, it is shown as the credibility of the information (such as website credibility, posting time, whether it is an original post or not and the authoritativeness of the poster etc.). The stage of dissemination is about the acceptance and identification of rumors and their impacts on the external environment. In Internet forum rumors, it can be manifested as the content of replies (such as number of posts, term frequency, parts of speech and sentiment etc.) and the characteristics of the corresponding securities market (such as stock indices, stock prices, fluctuations and turnovers, etc.).

Considering the characteristics of securities market rumors, this paper plans to take the five following characteristics: the features of the post (F_1), the features of replies (F_2), behavioral features of the poster (F_3), the feature of information credibility (F_4) and the features of the securities market (F_5). The specific content are as follows (as shown in Table 3):

Table 3. Feature Set Classification of Stock Forum Information of Eastmoney.com

Feature set	Feature	Content of Feature
F1	feature of post contents	features of term frequency, part of Speech and emotive words of the contents
F2	feature of replies	number of replies, term frequency feature, feature of part of speech and emotive words
F3	behavioral feature of poster	number of posts by the poster, number of followers and friends and duration of forum registry
F4	features of information credibility	credibility of the website, posting time, original posts or reposts, authoritativeness of the poster etc.
F5	feature of securities market corresponding to posting time	stock price change rate before and after the post, corresponding stock price change, change in volume and turnover rate etc.

Calculation of feature weight. The calculation of feature weight is involved in both text representation model and feature extraction, which is mainly reflects the importance of the feature item in the document representation. Salton put forward the TF-IDF algorithm (1975) and repeatedly demonstrated the effectiveness of the TF-IDF algorithm (1983) in information retrieval. Forman (2008) further measured the significance of categorical distribution with the statistical method. The main idea of TF-IDF is as follows: If a term appears in an article in high frequency but in low or rare frequency in other articles, then it can be said that the term has good differentiability and can be taken as a basis for classification. Based on this, a determination method of vocabulary weight (Q, Li et al., 2014a) is added in this paper to calculate the text feature weight according to the information feature sets (Figure 4) of the stock forum of eastmoney.com.

A weight is given in this paper to each term in the documents of rumor samples, the weight depending on frequency of the term in the documents of rumor samples, that is, the value of the weight is calculated as tf_{ij} according to the frequency of the term i in the document of rumor sample j.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

At the same time, if there is less documentation of the term i in the documents of rumor samples, meaning the term i had better differentiability of category, it is calculated as idf_{ij}, the idf will be greater.

$$idf_i = log\frac{\mathrm{N}}{df_i}$$

In order to deal with the phenomenon that may occur when the denominator df_i is 0 in the formula when i does not exist in the document collection of rumor samples, the formula above is amended as follows:

$$idf_i = log\frac{N}{df_i+1}$$

According to the idea of the TF-IDF model, the tf and idf of each term in each document of rumor samples are combined to form a comprehensive weight.

$$\mathrm{TF}-\mathrm{IDF}_{i,j} = tf_{i,j} \ \times idf_i$$

Thus, it can be concluded that if a term in a text of rumor sample appears in high frequency, but in lower frequency in other documents in the entire collection of rumor samples, then the value of the term's TF-IDF is high.

Text classification. As for the classification of Internet rumor text, the selection of classifier is the key and a main link. Text classifiers are generally divided into three types. One type is based on statistics, including Naive Bayes, K-Nearest-Neighbor (KNN) and Support Vector Machine. Based on supervised machine learning and text being represented by feature vector, this kind of classifier takes no language structure into account to obtain a generalized relationship. Another type is based on rules, mainly including the decision tree and correlation rules. This kind of classifier analyzes data sets, determines rules of classification, and then decides the category of unclassified text according to the rules. A third type is based on connection, mainly neural networks (NNET).

Yang and Liu (1999) proved by experiments that the classification performance of SVM is better than that of NNET or Naive Bayes, equivalent to that of KNN while the performance of the neural network classifier based on connection is not as good as that of SVM and KNN. As a result, SVM is taken as the classifier in this paper in the process of automatic identification of stock forum rumors of eastmoney.com.

SVM, proposed by Vapink (1999) based on the structural risk minimization (SRM) principle of statistics, is a method for supervised learning. SVM projects a vector to a higher dimensional space, and uses a subset of training examples to represent a decision boundary. The subset is called Support Vector. According to the structural risk minimization principle, the upper bound of generalizability will be increased with the increase in capacity of the model, and SVM can ensure minimum generalization errors under the worst condition. The JAVA program of LIBSVM-2.88, the most representative LIBSVM software package[1], is used in the paper to construct the SVM classifier. The main steps are as follows:

Step 1: determining the decision boundary. Screening rumors from the stock forum on eastmoney.com is actually a matter of binary classification containing N training samples. Each sample is represented as a binary *(x_i, y_i) $(i=1, 2, \ldots, N)$*, in which *$x_i=(x_{i1}, x_{i2}, \ldots, x_{id})^T$*, corresponding to the attribute set of sample *i*. The decision boundary of a linear classifier can be written as follows:

$$\mathrm{w} \cdot \mathrm{x} + \mathrm{b} = 0$$

where *w* and *b* are the parameters of the model.
The following method can be used to predict the category mark *y* of any test sample *z* as:

$$\mathrm{y} = \begin{cases} 1, & \mathrm{w} \cdot \mathrm{x} + \mathrm{b} > 0 \\ -1, & \mathrm{w} \cdot \mathrm{x} + \mathrm{b} < 0 \end{cases}$$

where *y=1* means rumor while *y=-1*, non-rumors.

Step 2: defining the boundary of classifier. By adjusting *w* and *b*, parameters of the decision boundary, the two parallel hyper planes on the boundaries between rumors and non-rumors, b_{i1} and b_{i2} can be represented as:

$$b_{i1:}\ \mathrm{w} \cdot \mathrm{x} + \mathrm{b} = 1$$
$$b_{i2:}\ \mathrm{w} \cdot \mathrm{x} + \mathrm{b} = -1$$

In order to calculate the boundary, suppose x_1 as a data point on b_{i1}, x_2 as a data point of b_{i2}, and put x_1 *and* x_2 respectively into the formulas of two hyper planes, the boundary *d* can be obtained by subtraction of the two formulas:

$$\mathrm{w} \cdot (\mathrm{x}_1 - \mathrm{x}_2) = 2$$
$$\|w\| \cdot \mathrm{d} = 2$$
$$\mathrm{d} = \frac{2}{\|w\|}$$

The polynomial kernel function is:

$$\mathrm{f(z)} = \mathrm{sign}(\textstyle\sum_{i=1}^{n} \lambda_i\ y_i\,(x_i\ \cdot z + 1)^2 + b)$$

Step 3: training and testing SVM Model. Compared with other classifiers, there is one more requirement in SVM, that is, the margin of the decision boundary must be the largest. The objective function is

$$\mathrm{f(w)} = \frac{\|w\|^2}{2}$$

The SVM learning tasks can be formally described as:

$$\min_{w} \frac{\|w\|^2}{2}$$

The above formula is subjected to *i=1, 2, …, N* (as shown in Figure 3)

[1] LIBSVM: an integrated SVM identification and regression software developed by Lin Chih-Jen, professor of National Taiwan University. With strong principles and high efficiency, the software is easy to use, providing open source codes, compiled files that can be executed in WINDOWS, and tested default parameters, so the users don't have to make much adjustment on the parameters of SVM algorithm design. The software also provides selections of kernel functions in common use, linear and polynomial, making it convenient to solve specific problems in SVM algorithm .

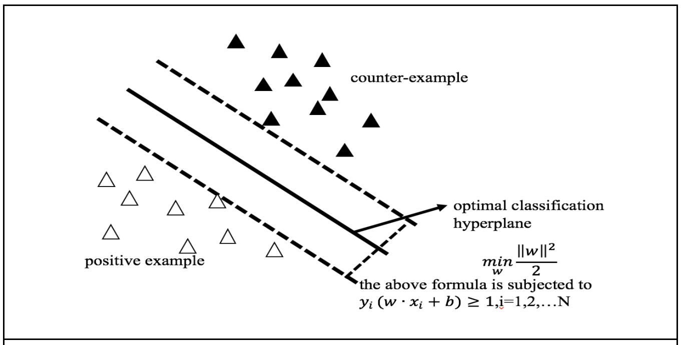

counter-example
optimal classification hyperplane
positive example
$\min_{w} \frac{\|w\|^2}{2}$
the above formula is subjected to
$y_i\,(w \cdot x_i + b) \geq 1$,i=1,2,…N

Figure 3. Diagram of SVM Classification

Step 4: evaluating SVM classifier. In order to test the accuracy of SVM classifier for Internet rumor classification, it needs to evaluate the classifier after SVM has completed sample testing so as to ensure the classification accuracy of all samples. At present, the evaluation indices of the classifier are roughly "*precision (P) + recall (R) + F-measure (F)*", algorithm precision, ROC and AUC, micro average and macro average etc. "*Precision (P) + recall (R) + F-measure (F)*" is used in the paper for evaluation. The specific formula is expressed as follows:

$$Precision(P) = \frac{a}{a+b}$$

$$Recall(R) = \frac{a}{a+c}$$

$$F - measure(F) = \frac{2\times\mathrm{P}\times\mathrm{R}}{P+R}$$

Where a is the number of classes of rumors correctly classified into rumors.
b is the number of non-rumors incorrectly classified into rumors.
c is the number of rumors incorrectly classified into non-rumors.

Step 5: determining sentimental polarities of Internet Rumors. After the forum information in the stock forum of eastmoney.com is screened, Internet rumors are automatically identified, thus significant information is obtained, such as the specific content of Internet rumors (text information), attention rates (amount of reading), information of rumor spreaders (their cyber IDs) and time for spreading rumors. However, little awareness is gained of sentimental polarities of the Internet rumors, or their sentimental direction (positive or negative). At this time, there comes the need to further define the sentimental polarities of screened Inter rumors so as to have a more comprehensive analysis about sentimental polarities and classify rumor spreaders' tendencies, opinions and attitudes, which can provide fundamental information to carry out next quantitative research. The technical diagram of determining sentimental polarities is as shown in Figure 4.

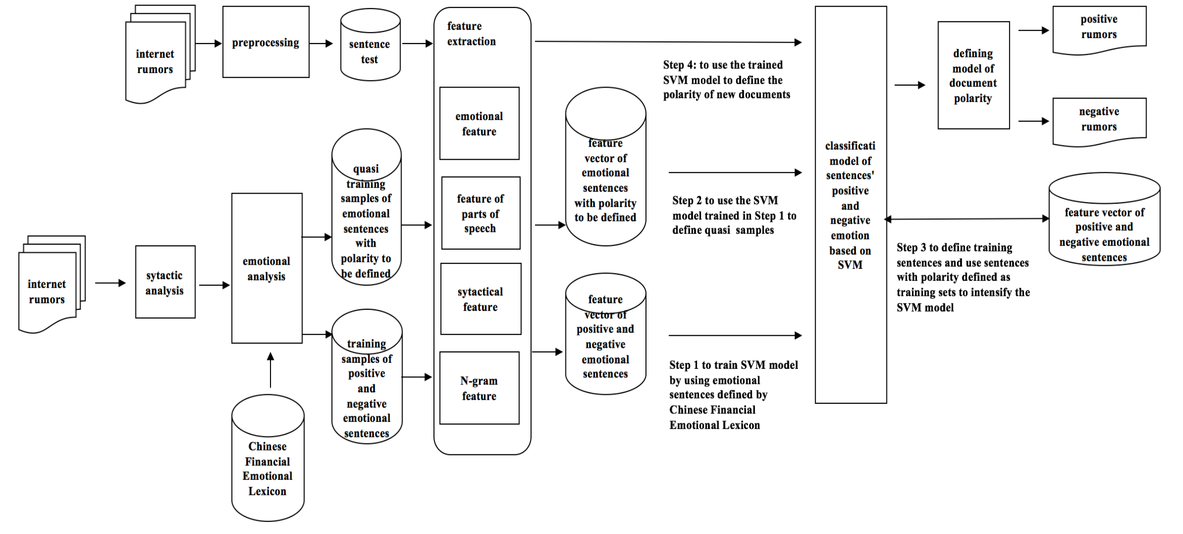

internet rumors
preprocessing
sentence test
feature extraction
emotional feature
feature of parts of speech
sytactical feature
N-gram feature
Step 4: to use the trained SVM model to define the polarity of new documents
feature vector of emotional sentences with polarity to be defined
Step 2 to use the SVM model trained in Step 1 to define quasi samples
feature vector of positive and negative emotional sentences
Step 1 to train SVM model by using emotional sentences defined by Chinese Financial Emotional Lexicon
classificati model of sentences' positive and negative emotion based on SVM
defining model of document polarity
positive rumors
negative rumors
feature vector of positive and negative emotional sentences
Step 3 to define training sentences and use sentences with polarity defined as training sets to intensify the SVM model
internet rumors
sytactic analysis
emotional analysis
quasi training samples of emotional sentences with polarity to be defined
training samples of positive and negative emotional sentences
Chinese Financial Emotional Lexicon

Figure 4. Technical Diagram of Defining Text Sentimental Polarities Based on Machine Learning

According to the Behavioral Financial Theory, the fluctuation of the securities market is subject to investor sentimental (De Long et al., 1990). In the quantization process of media texts, in addition to retaining proper nouns describing corporate fundamentals, sentimental words in media expression that affect investor psychology should be also taken into account. To extract the affective factors of texts, researchers have tried using sentimental analysis technique based on morphology or syntax to quantify the positive or negative sentimental tendency of a media document. For example, Tetlock et al. (2008) used the general sentimental lexicon Harvard-IV-4 to quantify the effects of affective factors in the news of securities market volatility by calculating the ratio of positive and negative words in the news. Schumaker et al. (2012) adopted OpinionFinder, a sentimental analysis software developed by the University of Pittsburgh to get the sentimental index of a piece of news, and they found that an integrated consideration of news nouns and sentimental index could more effectively help depict the connection of news and securities market volatility. But due to the lack of a Chinese financial sentimental lexicon, most researchers have to turn to artificial reading in order to improve the accuracy of sentimental analysis when doing research on the Chinese financial market, which greatly restricts the number of samples and adds up to the subjectivity differences in results (Peigong Li and Yifeng Shen, 2010; You, J.X and Wu, J., 2012).

The authors of the paper, based on the Chinese Financial Sentimental Lexicon (Q, Li et al.2014a), designed an unsupervised sentimental analysis algorithm for massive documents. The core ideas are as follows: (a) to use the Chinese Financial Sentimental Lexicon, plus syntactic parsing[2], to define the sentimental polarities of rumors and take the sentences as training sets whose tendencies shall be defined, and use the SVM to train a defining model oriented to the sentimental of sentences of rumors; (b) to derive sentences, which reach beyond a certain threshold value and are defined by the SVM as positive or negative from would-be-trained sentences, as new sentences to be trained so as to improve the learning ability of SVM; (c) for a new piece of rumor information, the well-trained SVM will define its positive or negative polarity and the sentimental polarities of the entire piece of rumor information will be defined according to the sentences' sentimental polarities in the text and their importance in the position of the text[3]. The merit of this diagram lies in using machine learning algorithms to overcome the low recall rate of sentimental defining only relying on the sentimental lexicon, and avoid time-consuming large-scale manual construction of training samples. Thus, the diagram is a fit for processing large quantities of text. The specific classification principles based on SVM's definition of sentimental polarities are as previously mentioned.

Experiment

Experimental Method and Data

Experimental Method

The method of 10-fold cross-validation is used to evaluate the experiment. With this method, the text information of the test set (K) chosen is divided at random into 10 segments, of which 9 segments are used for training and the other segment for test. The process is repeated 10 times, using different test text each time. The specific steps are as follows:
Step 1: 1 segment is retained in each iteration.
Step 2: the other 9 segments of text information are used for training the classifier.
Step 3: the retained text information is used to test the classifier and save the test results.
The above 3 steps are repeated 10 times.

Experimental Data

Step 1: Sample Selection： This paper build the "*Lexicon of Chinese Securities Market Rumors"* by using the rumor clarification announcements of China Securities Regulatory Commission. From which we selected 10 thousand rumors as information samples (A) and 10 thousand non-rumors as information samples(B) both by random. Thus, a total of 20 thousand pieces of text information (M=A+B) are taken as the data samples for experiment.

[2] Related rules are made according to the rules of syntactic parsing analysis which can thus solve the interference of comparison, transition, negation and other sentence patterns in defining sentimental tendency.

[3] The title of the article, the initial sentence, the first sentence of paragraphs and sentences in paragraphs vary in sentimental importance.

Step 2: Building Training Sets (H=C+D): 90% of rumor samples (C) and 90% of non-rumor samples (D) and are randomly selected from the sample data and are evenly divided into 10 segments as the training sets.

Step 3: Building Test Sets (K=E+F): The remaining 10% of rumor samples (E=A-C) and 10% of non-rumor samples (F=B-D) are evenly divided into 10 segment as the test sets.

Experimental Evaluation

According to the prescribed method, the prepared experimental data and "10-fold cross-validation" are used to conduct evaluation on the experiment, the results of which are as follows(as shown in Table 4):

Table 4. Results of 10-Fold Cross Validation

Times of Validation	Sample			Classified as Rumor			Classified as Non-rumor
	Actual category	Polarity	Quantity	Subtotal	Quantity	Sentiment Accuracy	Quantity
Total	Rumor Info.	Positive	7000	7232	5166	74%	2768
		Negative	3000		2066	69%	
	Non-rumor Info.	—	10000	2181	—		7819

(Remarks: The results of each of the ten times of validation are omitted here.)

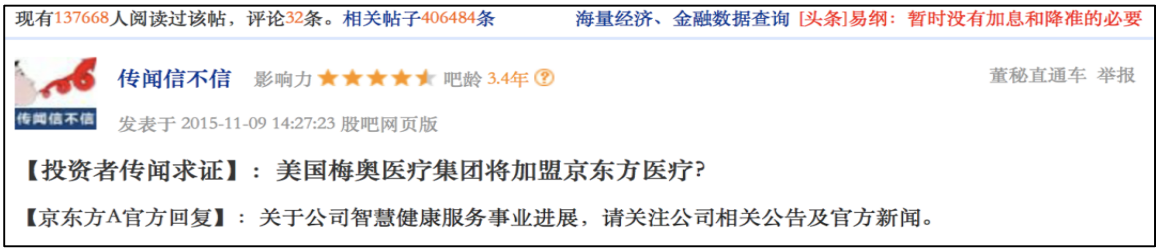
Evaluation Indices of SVM Classifier Performance:

$$Precision: \ P = \frac{\mathrm{a}}{\mathrm{a+b}} = \frac{7232}{7232+2181} = 76.82\%$$

$$Recall: \ R = \frac{\mathrm{a}}{\mathrm{a+c}} = \frac{7232}{7232+2768} = 72.32\%$$

$$F - measure: \ F = \frac{2\times \mathrm{P} \times \mathrm{R}}{P+R} = \frac{2\times 76.82\% \times 72.32\%}{76.82\% + 72.32\%} = 74.50\%$$

According to the above-mentioned "Results of Ten-Fold-Cross Validation" and "Evaluation Indices of SVM Classifier Performance", the authors of the paper hold that: (1) the comprehensive classification rate of SVM classifier reaches 74.50%, which suggests that the classifier performs well in classifying Internet rumors and can automatically identify the "Internet rumors" crawled by the web crawler from all the "stock forum" of eastmoney.com; (2) the average accuracy rate of classification of sentimental polarities of Internet rumor reaches 71.5%, which suggests that the classifier performs well in classifying sentimental polarities (Q. Li, et al., 2014a), and the well-trained SVM classifier can classify the sentimental polarities of Internet rumors; (3) methods such as the decision tree and Naive Bayes are used for automatic identification, but with less desirable effects than SVM, which proves the reliability of the study in this paper. Finally, the well-trained SVM classifier automatically identified about 430 thousand rumors from 37 million 600 thousand pieces of information (screen-shot of rumors is as shown Figure 5)



现有137668人阅读过该帖，评论32条。相关帖子406484条　海量经济、金融数据查询 [头条]易纲：暂时没有加息和降准的必要

发表于 2015-11-09 14:27:23 股吧网页版

【投资者传闻求证】：美国梅奥医疗集团将加盟京东方医疗?

【京东方A官方回复】：关于公司智慧健康服务事业进展，请关注公司相关公告及官方新闻。

Figure 5. Screen-shot of Rumors in the Eastmoney "Stock Forum"

Conclusion

The paper has successfully crawled around 40 million massive data from the stock forum of eastmoney.com, solving the difficulty of crawling massive Internet forum text information. By using the machine learning method (SVM), the authors put forward a research method for automatic identification of Internet forum rumors for the first time and achieved good comprehensive classification through experimentation and evaluation, and proved that the SVM classifier with reliable performance is able to realize automatic identification of rumor information in the stock forum of eastmoney.com. This is a constructive attempt for further exploration of in-depth integration of machine learning and financial industry and provides substantial data and experiences for follow-up study on the incidence relationship between Internet rumors and the securities market. The main contribution of this paper lies in the successful crawling of massive Internet forum data with computer technologies, automatic identification of rumors by the use of intelligent technology, which is no longer confined to case study or statistical sampling of rumors and has laid a solid foundation for more profound and scientific research on the effects of Internet media on the securities market. This paper also describes the current situation and characteristics of the “Internet rumors” in China’s securities market for the first time, providing an important reference for the participants in the securities market, which is of great practical significance.

Acknowledgments

This work has been supported by the Fundamental Research Funds for the Central Universities (Nos. JBK1507K04).

References

Antweiler, W., Frank, M, Z., (2004). Is all that just noise? The information content of Internet stock message boards, Journal of Finance, pp. 1259-1294.

B. M. Barber and D. Loeffler, (1993). The “dartboard column: Second-hand information and price pressure”, Journal of Financial and Quantitative Analysis, vol. 28, no. 02, pp.273–284.

Bollen, J., Mao, H., Zeng, X. J., (2011). Twitter mood predicts the stock market, Journal of Computation Science, 2(1), pp. 1-8.

Curme, C., Preis, T., Stanley, H. E., Moat, H. S., (2014). Quantifying the semantics of search behavior before stock market moves, Proceedings of the National Academy of Sciences (PNAS), 111(32), pp. 11600-11605.

De long J B, Shleifer A, Waldmann R, (1990). Noise Trader Risk in Financial Markets, Journal of Political Economy, 98(4): pp. 703-738.

Diefenback, (1972). How good is institutional barkerage research, Financial analyst journal, V28, pp. 54-60

DiFonzo N, Bordia P, Rosnow RL, (1994). Reining in rumors. Organizational Dynamic, 23(1): pp. 47-62

Forman G, (2008). Bns feature scaling: an improved representation over TF-IDF for SVM text classification, Proceedings of the 17th ACM Conference on Information and Knowledge Management. USA，California: ACM, pp. 263-270.

G Salton, A, Wong and C. S. Yang Cornell, (1974). A vector space model for automatic indexing, Communications of Acm, 18(11): pp.613–620.

G Salton，CT Yu, (1975). On the construction of effective vocabularies for information retrieval, ACM Sigplan Notices，9(3) : pp. 48-60.

H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis, (2013). “Quantifying wikipedia usage patterns before stock market moves,” Scientific reports, vol. 3, pp. 1–5.

I. Mathur and A. Waheed, (1995). “Stock price reactions to securities recommended in business week’s inside wall street,” Financial Review, vol. 30, no. 3, pp. 583–604.

J Pound, R Zeckhauser, (1990). Clearly Heard on the Street: The Effect of Takeover Rumors on Stock Prices, Journal of Business,63(3): pp. 291-308.

Kapferer, Ruolin Zheng Translate, (2008). Rumors - the world's oldest media, Shanghai People's Publishing House.

Lavrenko, V., Schmill, M., Laurie, D., Ogilvie, P., Jensen, D., Allan, J, (2000). Language models for financial news recommendation, IN: Proceedings of the 9th International conference on information and Knowledge Management(Cikm), pp. 389-396.

Peigong Li, Yifeng Shen, (2010). The Corporate Governance Role of Media：Empirical Evidence from China，economic Research，Vol. 4, pp. 14-27.

M.A. Mittermayer and G. F. Knolmayer, (2006). "Newscats: A news categorization and trading system." IEEE, pp. 1002–1007.

Miller G, Beckwith R, (1990). Introduction to Wordnet: An Online Lexical Database, International Journal of Lexicography,3(4): pp. 234-244.

P. Clarkson, D. Joyce, and I. Tutticci, (2006). "Market reaction to takeover rumor in internet discussion sites," Accounting and Finance, vol. 46, no. 1, pp. 31–52.

P. L. Davies and M. Canes, (1978). "Stock prices and the publication of second-hand information," Journal of Business, vol. 51, no. 1, pp. 43–56.

Q. Li., Wang, T., Gong, Q., Chen, Y., Lin, Z., Song, S. k., (2014a). Media-aware quantitative trading based on public Web information, Decision Support Systems 61, pp. 93-105.

Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, (2014b). "The effect of news and public mood on stock movements," Information Sciences, vol. 278, pp. 826–840.

Q. Li., Lin, J. J., Li, P., Chen. H, (2015). tensor-Based Learning for Predicting Stock Movements, In: Proceedings of the 29th AAAI Conference on Artificial Intelligence.

R. P. Schumaker, Y. Zhang, C.-N. Huang, H. Chen, (2012). "Evaluating sentiment in financial news articles," Decision Support Systems, vol. 53, no. 3, pp. 458–464.

Rose，(1951). Rumor in the Stock Market，Public Opinion Quarterly, 15(3):pp. 461-486.

Salton G. (1983). Extended boolean information retrieval, Cornell University，11(4): pp.95-98.

Siganos, A., Vagenas-Nanos, E., Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets, Journal of economic Behavior & Organization 107, 730-743.

Spiegel, U., Tavor, T., Templeman, J. (2010). The effects of rumors on financial market efficiency, Applied Economics Letters, 17(15), pp. 1461-1464.

Sui. Y. P., (2015). Empirical Research on the Impact of Rumors on The Stock Prices, Harbin Institute of Technology.

T. Preis, H. S. Moat, and H. E. Stanley, (2013). "Quantifying trading behavior in financial markets using google trends," Scientific reports, vol. 3, pp. 1–6.

Tetlock, p. C., Saar-Tsechansky, M., Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals, Journal of Finance, 63(3), pp. 1437-1467.

Vapnik V N, (1998). An overview of statistical learning theory, Neural Networks IEEE Transactions on 10(5), pp. 988-999.

W. S. Chan, (2003). "Stock price reaction to news and no-news: Drift and reversal after headlines," Journal of Financial Economics, vol. 70, no. 2, pp. 223–260.

Wang. B., Huang. H., Wang. X, (2011). A novel text mining approach to financial time series forecasting, Neuro computing 83, pp.136-145.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, j., lam, W, (1998). Daily stock market forecast from textual Web data, In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp. 2720-2725.

X. Luo, J. Zhang, and W. Duan, (2013). "Social media and firm equity value," Information Systems Research, vol. 24, no. 1, pp. 146–163.

You. J. X., Wu. J., (2012). Spiral of Silence：Media Sentiment and the Asset Mispricing, Economic Research, Vol.7，pp. 141-151.

Y. Yang, X. Liu A, (1999). re-examination of text categorization methods, In: Proceedings of sigir-99, 22nd ACM International Conference on Research & Development in Information Retrieval, pp. 42-49.

Zhao. J. M., He. X., Wu. F. Y., (2010). Study on stock market rumors of China: Spreading and clarification of the rumors and the effect to the stock price, Management World, Vol.11: pp. 38-51.