Internet Financial News and Prediction for Stock Market: An Empirical Analysis of Tourism Plate Based on LDA and SVM

Jinxiao Wang
Tsinghua University, Beijing, China
Email: wangjx.18@sem.tsinghua.edu.cn

Jiaxin Shi
Peking University HSBC Business School, Shenzhen, China
Email: shijiaxin1997@foxmail.com

Dexin Han and Xiaoyu Zhao
China University of Political Science and Law, Beijing, China
Email: hdxjasmine@sina.com, cupl_zxy@163.com

***Abstract*—Internet financial news plays an important role in stock market forecasting. This paper discusses the relationship between the content of the Internet financial news and the yield of the stock market by using text mining technology and machine learning technology. The Latent Dirichlet distribution (LDA) model is used to analyze the Internet financial news. And the support vector machine (SVM) algorithm is used to predict the trend of the sector. Afterward constructs a trading strategy. The results show that the introduction of the information of tourism topic distribution in the Internet financial news can effectively improve the accuracy rate of forecast, thus increasing return of investment, especially when the stock market is in a volatile period. To sum up, the information of Internet.**

***Index Terms*—internet financial news, stock market forecast, text mining, support vector machine**

I. Introduction

Nowadays, the Internet has become the main source of information for public to access, especially the Internet Finance Module, which has become an indispensable way for investors to obtain market information [1]. In this context, the extraction and mining of internet financial news is of great significance for discovering market conditions.

This paper takes the tourism sector as the research object, and obtains more than 80,000 financial news from November 16th, 2011 to July 11th, 2015 by text mining technology, which is on the financial news column of Phoenix Finance website. The Latent Dirichlet distribution (LDA) model is used to analyze the Internet financial news in depth, and then combined with the historical information of the stock market, the support vector machine (SVM) algorithm is used to predict the trend of the sector, and finally the trading strategy is constructed.

Compared with the existing research on the relationship between Financial News and stock price prediction [2], this paper has the following sparking points. Firstly, we study on the tourism sector specifically, and combine the historical information of the stock market with the news information related to the tourism sector on the financial and economic websites to build the prediction model. Secondly, by comparing the changes in the accuracy of the prediction model before and after the introduction of Internet financial news information, the role of Internet financial news can be objectively demonstrated [3]. Thirdly, the data span is longer, which improves the shortcomings of the existing literature research. Fourthly, we conducted a detailed study of different stock market stages (up, down, and volatility), respectively exploring the role of Internet financial news in predicting the trend of the stock market in different stages.

II. Introduction of the Model

A. Latent Dirichlet Distribution (LDA) Model

This paper selects the Latent Dirichlet distribution (LDA) model as the extraction method of the hot topics in Internet financial news. As a probabilistic generation model, the LDA model can map high-dimensional feature vectors to low-dimensional semantic space. Since text is composed of different topics and topics are the main ideas composed of different words, the LDA model can effectively identify the topic information contained in large-scale documents [4].

Fig. 1 shows the LDA model diagram, where the solid point represents implicit variables such as the distribution of words in the topic model, the hollow points represent implicit variables such as topic distribution parameters in

Manuscript received February 11, 2019; revised July 23, 2019.

doi: 10.12720/jait.10.3.95-99

the model, and the rectangle represents the process of repeated sampling of the document. The outer rectangle represents the corpus, and the inner rectangle represents the repeated sampling of the subject and words for each document. The relevant symbols are defined as follows:

1) For a text, the basic data unit is the feature item, and here is the word of the text, with the item {1, …, V} representing the vocabulary. The vth word in the vocabulary can be expressed as a V-dimensional vector.

2) Use $W = \{w_1, w_2, w_3, \ldots\ldots, w_n\}$ to represent a document, w_n represents the nth feature word in the document.

3) Use D to represent a collection which contains M texts, ie a corpus; a text set D can be represented as $D = \{W_1, W_2, \ldots, W_n\}$.

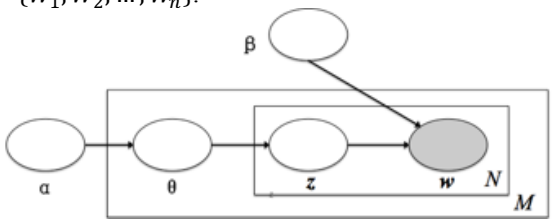

Figure 1. LDA model diagram

The premise of classifying text using LDA is the determination of the distribution of implicit variables, that is, the process of generating a document by the implicit subjects in the text. In the LDA model, the process of generating each document M is as follows:

1) First, to get the number of words in a document, the process is implemented by Poisson(ξ) $(N \sim Poisson(\xi))$.

2) Calculate the probability distribution vector of the topic for each piece of text using the Dirichlet distribution $(\theta \sim Dir(\alpha))$.

3) For each word w_n in N:

a) Select a topic item $Z_n \sim Mutinormial(\theta)$ from the topic distribution;

b) Select w_n from a ditional probability distribution $p(w_n|z_n, \beta)$.

Then give the parameters α, β; you can get the joint distribution of an article as follows:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(w_n|z_n, \beta) \quad (1)$$

By iterating θ and summing up z, we can get the edge probability distribution of an article:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta))d\theta$$

Finally, based on the edge probability distribution of each article, the joint probability distribution of the entire corpus can be obtained:

$$p(D|\alpha, \beta) =$$

$$\prod_{n=1}^{N} \int p(\theta_d|\alpha) \sum_{z_{d_n}} p(z_{d_n}|\theta_d) p(w_{d_n}|z_{d_n}, \beta) d\theta_d \quad (2)$$

The solution of the model is obtained by Gibbs Sampling's method to get the posterior distribution of the topic distribution and word distribution to determine the parameters α and β.

B. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a data mining technology based on statistical learning theory, which is essentially a binary classification model. It aims to maximize the distance between categories and automatically find the support vector with the strongest ability to distinguish categories.

Schematic diagram of support vector machine segmentation hyperplane.

Suppose the training set of the sample is $X = \{x_1, x_2, \ldots, x_n\}, X \in R^d$. The corresponding mark of training set X is training $\{y_1, y_2, \ldots, y_n\}, y_i \in \{1, -1\}d$, which is the dimension of the training set sample space. Now, we need to find a discriminant function g(x)=$\omega \cdot x + b$ to make $g(x_i) \in \{-1,1\}$ for any x in X, and the classification interval can be described as $2/||\omega||$. If you want the maximum interval between categories, the value of w should be the smallest. The problem above can be seen as the optimization of the following equation:

$$min||\omega||/2 \quad (3)$$

$$s.t.\, y_i[\omega x_i + b] - 1 \geq 0, i = 1,2,\ldots,n \quad (4)$$

The above two expressions are convex functions, so the optimization problem of SVM is to solve the above-mentioned quadratic convex optimization. Then, in the two-class problem, the global optimal solution of the above quadratic programming is the solution of the SVM. With Lagrange multiplier optimization, there are:

$$f(x) = sgn(\omega x * + b *) \quad (5)$$

In the above formula, a* and b* are the classification hyperplane parameters, (x_i x)represents the vector product of the two vectors. For nonlinear problems, the SVM is processed by transforming the nonlinear problem into a linear problem by the change of the kernel function, so that the SVM can map the low-dimensional space corresponding to the nonlinear problem to the high-dimensional space corresponding to the linear problem. Then, in high-dimensional space, nonlinear problems are transformed into linear separable. The problem at this point can be transformed into the following form:

$$f(x) = sgn\left(\sum_1^k a * y_i K(x_i x) + b *\right) \quad (6)$$

III. Source and Processing of Data

A. Financial News Text Source and Pretreatment

The research object of this paper is Internet Finance News. And we use text mining technology to convert a large amount of unstructured text into structured data that can be processed by computers. This paper mainly uses Python to get 80,000 financial news from the Phoenix Financial website's securities news column, the time span is from November 16, 2011 to July 11, 2015. Some special noise URLs were also processed during the crawl.

B. Extraction of Web Page Text Information

On the basis of crawling financial news, it is necessary to preprocess the text to effectively extract the information. This paper focuses on three key processes and techniques.

1) Text segmentation

Compared to a single word, phrases include more complete semantic information, which can express the content of the text more accurately. Therefore, we use the steps of word segmentation to extract valid information from Chinese text. The word segmentation system used in this paper is ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System), which is the best system for Chinese word segmentation. This system includes the functions of multiple different modules such as named entity recognition, Chinese word segmentation and part-of-speech tagging. This paper will mainly use the participles and part-of-speech annotations functions, the final selection is some practical verbs, nouns, adjectives, quantifiers and so on.

2) Feature expression and key techniques for dimensionality reduction

After the word segmentation, as the structure of the text is more complicated, the dimension of the obtained word collection is very high and the obtained word collection cannot be directly extracted from the feature. Therefore, it is necessary to extract as few features as possible from the text to represent its content. The feature dimension reduction method selected in this paper is TF-IDF (term frequency–inverse document frequency), which is based on the document frequency. TF-IDF not only has a high degree of accuracy, but also combines weighting the importance of a feature.

3) Hot topic recognition of financial news

Enter news texts that have undergone text vectorization and dimensionality reduction. This paper uses LDA to output the appearing probability of each text under each theme and the corresponding high-weight keywords, and extract hot topics from financial news. The results show that there is a clearer meaning of a set of topics. Below Table I and Table II are some of the keywords that correspond to the travel theme and the probability of the daily occurrence of the topic i_t:

TABLE I. KEYWORDS CORRESPONDING TO THE TRAVEL THEME

Travel	Tourist	Scenic spot	Tourism	Travel agency	Ticket	Vacation	person-time

TABLE II. THE PROBABILITY OF THE DAILY APPEARANCE OF THE TRAVEL THEME

t	2011.11.16	2011.11.17	…	2015.07.10	2015.07.11
i_t	0	0.0110643	…	0.0240629	0.064512

C. Stock Data Source and Pretreatment

This paper selects the Shanghai and Shenzhen 300 Index to reflect the overall situation of the stock market. Market yield is r_m , $r_m = 100 \times (In\ p_t - \ln p_{t-1}), p_t$, represents the closing price of the Shanghai-Shenzhen 300 Index on the t-day. The tourism sector yield is represented by r, the tourism sector yield is $r_b, r_b = 100 \times (\ln p'_t - \ln p_{t-1}')$, p_t' represents the closing price of the t-day tourism sector index.

IV. EMPIRICAL ANALYSIS

First, we select the data from November 16, 2011 to July 11, 2015. According to the ups and downs between the closing price of the tourism sector index and the closing price of the day before, that is, the positive and negative of r_b, the rise of the tourism sector (denoted as 1) and the decline (denoted as -1).We divide the data into two parts: 70% of the data in trading day is classified as a training set, and the remaining 30% is a forecast set. The rising and falling of the tourism sector is used as a classification label, and the previous day's CSI 300 index yield is used as the classification basis. The prediction accuracy rate of the SVM model is 50.9506%. After the probability information of the topic is added as the basis of judgment, the discriminant accuracy is increased to 54.5113%.

In addition, the paper also divides the stock data into three segments according to the trend of the market for detailed research. During the period from December 5, 2012 to February 8, 2013, the overall trend of the market rose. At this stage, the prediction accuracy rate is as high as 81.6092%. Judging from the relevant information of Internet financial news, the forecast accuracy rate has risen slightly to 82.7586%.

Fig. 2 shows the results of the discriminant analysis at this stage. The classification label for the blue sample points is rising (denoted as 1), and the label for the red sample points is falling (denoted as -1). The abscissa of the sample is the CSI 300 yield after standardization on the day before. The ordinate of the sample is the probability of traveling related topics in the Phoenix Finance website. It can be seen that when the whole stock market is rising, the plate yield and the large-cap yield are closely linked. The introduction of news information has a certain improvement effect on the stock market forecast.

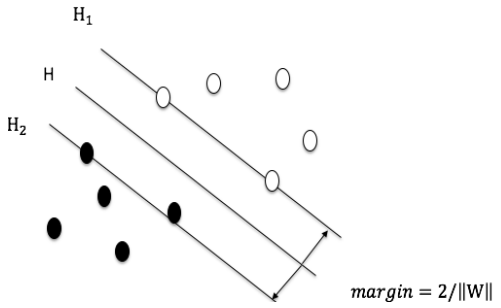

Figure 2. SVM model diagram

Due to the sample stage of the study, the decline period of China's overall stock market has been very short. In order to increase the accuracy of the forecast, we will splice three stages of decline from November 16, 2011 to January 6, 2012, May 7, 2012 to September 17, 2012, and June 17, 2015 to July 13th, 2015 as a large drop to expand the sample size. The empirical results show that in the overall decline of the market, just the previous day's CSI 300 yield rate for the disciplinary analysis of

the tourism sector, the accuracy rate can reach 50.7937%. If the Internet financial news information is used for identification, the accuracy can be improved to 55.5556%.

This paper also selects the volatility section for research, from March 17, 2014 to July 30, 2014. The result shows that when the Internet financial news information was not included for forecasting, the prediction accuracy rate was as low as 41.3793%. After adding the information, the accuracy of the evaluation of the tourism sector rose sharply to 62.069 %.

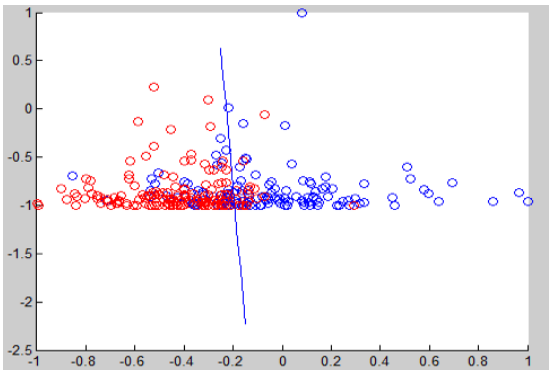

Figure 3. Discrimination results of SVM model in the rising segment

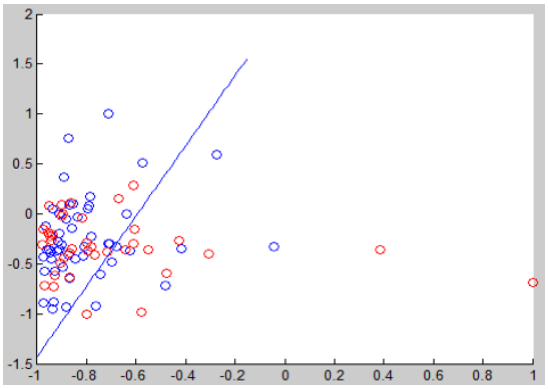

Figure 4. Discrimination results of SVM model in the fluctuating segment

In general, after adding the relevant topic probability information of the tourism sector in the Internet financial news on the day, the forecast accuracy rate of the day's tourism sector has increased. The results of the segmentation forecast show that in the bull market or bear market stage, the previous day's ups and downs of the market can provide more prediction basis for the day's yield, and the accuracy can be improved slightly after adding news information. The stock market is relatively volatile in fluctuating segment due to the trend of the market is not clear. The ups and downs of the sector cannot rely on the information of the previous day's market earnings to make effective predictions. At this time, the introduction of Internet financial news information can greatly improve the forecasting effect.

From the training SVM model (as shown in Fig. 3 and Fig. 4), it can be seen that the sample points of plate uptrend (shown as blue in the figure) are mostly if there are many Internet financial news related to tourism themes on that day, the tourism sector tends to rise. According to this phenomenon, we can make further study to learn the impact of Internet financial information on the stock market mechanism.

After joining the Internet financial news information, the SVM prediction model proposed by us reaches a correct rate of over 55% to forecast the intraday plate trend. This result is a very meaningful result for the people that believes in the law of large numbers in quantitative trading field. The higher correct rate of prediction can bring considerable profits, and the investment strategy can be further constructed on this basis.

V. Investment Strategy Construction

A. Investment Strategy

One of the most important purposes of studying the stock market is to study trading strategies. When there is a positive rate of return, the buying transaction is executed; when the yield is lower than expected, it is not traded or closed, reducing economic losses.

TABLE III. Investment Strategy-Trading Strategy and Rate of Return

$r_{m(t-1)}$	Forecast ups and downs	r_{it}	Actual rise and fall	Predictive error	Trading strategy	Rate of return
0.733695453	1	1.925730679	1	Right	Trading	1.925730679
1.094704232	1	1.774594813	1	Right	Trading	1.774594813
-1.012118901	-1	0.905829905	1	Wrong	Not trading	0
0.240511342	1	0.818768832	1	Right	Trading	0.818768832
1.054598482	1	0.183075652	1	Right	Trading	0.183075652
-0.820520792	-1	1.807325755	1	Wrong	Not trading	0
...	...	...	...	...	...	...
1.022565859	1	-1.460690655	-1	Wrong	Trading	-1.460690655
			Accuracy	0.413793103	Total return	6.131113645

Investment Strategy 1: This paper studies whether to buy at the close of the previous day and sell it at the close of the day. The transaction costs such as transaction costs are not considered here. The construction of investment strategy 1 is only based only on the historical revenue information of the market. If the closing price of the CSI 300 Index on Tuesday is higher than that on Monday, it is considered that the market's yield on Tuesday is greater than 0. Therefore, the investors will buy the portfolio of the tourism sector at the close of Tuesday and sell it at the close of Wednesday, and do not trade on the contrary. The data from 64 trading days from March 2014 to July 2014 were selected as training samples, and 29 data in 2014 were used as prediction samples. The final result is shown in Table III.

TABLE IV. Trading Strategy and Rate of Return of Investment Strategy2

$r_{m(t-1)}$	i_t	Predicted result	r_{it}	Actual rise and fall	Predictive error	Trading strategy	Rate of return
0.733695453	0.085300712	1	1.925730679	1	Right	Trading	1.925730679
1.094704232	0.075628773	1	1.774594813	1	Right	Not trading	1.774594813
-1.012118901	0.271399833	-1	0.905829905	1	Wrong	Not trading	0
0.947056691	0.032629151	1	0.943075289	1	Right	Trading	0.943075289
-0.068366032	0.072810205	1	-0.646566254	-1	Wrong	Trading	-0.646566254
...	...	...	...	...	...	...	...
-1.310818153	0	1	0.009870633	1	Right	Trading	0.009870633
				Accuracy	0.62068966	Total return	12.11942065

Investment Strategy 2: Use the SVM model built with the Internet financial news information to predict the ups and downs. If the forecast result is up, choose to trade, otherwise don't trade. Table IV is the trading strategy and rate of return after adopting investment strategy 2.

B. Validation of Investment Strategy

It can be seen from Table IV that the classification accuracy of the prediction set by the model obtained after training reaches 62.069%, which is significantly better than the accuracy of 41.379% based solely on the historical information of market return. In the forecast period, only based on the historical information of the market rate of return, the cumulative rate of return can be 6.13%, and according to the investment strategy 2, the cumulative rate of return that can be obtained is 12.12%. Besides, after making sector investment by this strategy, the rate of return is 9.41% higher than the year-on-year increase of the Shanghai index in the same year, which indicates that the Internet financial news has great application value for constructing investment strategy.

VI. Conclusion

With the economic development brought about by reform and opening up, people's investment awareness has gradually changed, and stocks have become an important part of Chinese investment and financial management [5]. At the same time, the influence of the news media in the stock market is also growing. In this context, quantifying text information to analyze the stock market has important theoretical and practical value.

This paper innovatively constructed a model to predict the ups and downs of the sector, mainly taking into account the historical information of the stock market. And we can find that the Internet financial news has a significant effect on improving the accuracy of the forecasting model. Besides, this paper also discusses the role of Internet financial news in different stages on the forecast of the trend of the sectors. The research results of this paper show that whether it is the whole or the segmentation study, the information about the distribution of the relevant topics of the tourism sector in the Internet financial news on the day has improved the accuracy of forecasting the ups and downs of the tourism sector. Especially when the stock market is in a period of volatility, the information related to financial news of the sector can greatly improve the accuracy of the forecast. Finally, this paper constructs an investment strategy based on the prediction model of the Internet financial news topic.

The future research can further study the mechanism and path of Internet financial news influencing stock market based on the existing research technology [6], [7]. At the same time, this paper only mines qualitative news ontology information, and there are still many factors affecting the trend of stock price, so this paper hopes to expand the scope of text mining and analysis in the future research, so as to improve the accuracy of prediction model.

References

[1] H. Liu and J. Xu, "The impact of internet heterogeneous financial news on the stock market: Evidence from Chinese internet data and listed companies," *Industrial Economic Research*, no. 1, pp. 76-88, 2017.

[2] J. Yang, "An empirical analysis of the impact of internet financial news on stocks," Southwestern University of Finance and Economics, 2012.

[3] L. Zhao, Q. Zhao, J. Yang, T. Wang, and Q. Li, "Quantitative analysis of the impact of financial news on china's stock market," *Journal of Shandong University (Science Edition)*, no. 7, pp. 70-75, 80, 2012.

[4] M. A. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *Proc. Hawaii's International Conference on System Sciences*, 2004, p. 10.

[5] Q. Zhao, "Research on the impact of internet financial news on china's stock market," Southwestern University of Finance and Economics, 2012.

[6] X. Kong, X. Bi, and S. Zhang, "Financial news and stock market forecast——An empirical analysis based on data mining technology," *Mathematical Statistics and Management*, no. 2, pp. 215-224, 2016.

[7] X. Meng, Y. Yang, and X. Zhao, "Financial news and stock market investment strategy research——Text mining based on financial website," *Research on Investment*, no. 8, pp. 29-37, 2016.

Jinxiao Wang received her B.S. degree in applied mathematics and economics from China University of Political Science and Law in 2014. She is a Ph.D. candidate of School of Economics and Management, Tsinghua University now. Her research interests include technology innovation, entrepreneurship and strategy management.

Jiaxin Shi received his B.S. degree in vehicle engineering from Tsinghua University in 2014. He is a postgraduate student of Peking University HSBC Business School now. His research interests include technological innovation performance and financial management.

Dexin Han is an undergraduate student of School of Business, China University of Political Science and Law. She majors in business management. She has a strong interest in academic research on finance and Information Technology.

Xiaoyu Zhao is an undergraduate student of School of Business, China University of Political Science and Law. And she will receive her B.S. degree in applied mathematics and economics from China University of Political Science and Law in 2020. Her research interests include finance, industrial economics and technical economics.