Massive Text Mining for Abnormal Market Trend Detection

Ying Li*†, Ting Jin*, Meng Xi*, Shengpeng Liu*, Zhiling Luo*‡
*College of Computer Science, Zhejiang University, Hangzhou, China
†Binhai Industrial Technology Research Institute of ZheJiang Univerity, Tianjin, China
‡Corresponding Author, Email: luozhiling@zju.edu.cn

***Abstract*—The sentiment behind financial text has been observed to have correlations with stock market trend. Though widely discussed, the study on this topic faces the challenge coming from the lack of open dataset and labeled financial text. In this work, we collected a large amount of Chinese financial text from financial news, research report, stock BBS and corporate announcements. It contains 3 million articles about 128 stocks from 2010 to 2018. And then we proposed a model mapping from the text and latent sentiment to the abnormal market trend. It combines the posting amount, daily market index with the RBM-embedded document vector, and extracts the abnormal features via LSTM. After that a neural net is employed to identify the abnormal trend. The experimental results on our dataset show the effectiveness of our approach comparing to baseline methods.**

***Index Terms*—Market Trend, Financial Text, RBM, LSTM**

I. Introduction

With the rapid growth of Internet and the prosperity of the stock market, information technology has brought about significant changes in the development of the financial industry [1]. A growing number of institutions and individuals tends to publish financial information and express their mood-states or emotion on the Internet. Information dissemination by many authoritative organizations or experts as well as coverage of some important financial incidents has become the benchmark for the development and changes in some industries or fields. Therefore, it is of great importance to analyze and summarize huge financial information in time so as to obtain useful knowledge, which has attracted more and more attention from academia as well as business [2].

Scores of traders always get caught in such common scenario that the market moves in one straight line without pull back in a period of time and while it begins to change direction from its original trends gradually, the fear of losing initial gain makes most traders still maintain their attitude to the original trend. The above scenario usually makes traders give back all the gain and fall into confusions. Therefore, it is of great significance to find out abnormal market trend of the stock in time for the investors. However, there exists massive stock textual data presented in front of investors, which makes them easy to fall into confusion. Then, some potential but essential information is not taken into consideration, resulting in unnecessary losses. Thus, it is meaningful for traders and data analysts to be able to analyze massive data timely, to filter out some meaningless information, and to extract pivotal information.

There are two main challenges in the study of abnormal market trends detection. First of all, there is no open annoted dataset available for Chinese financial text mining. We have to collect financial text and do adata preprocessing. An abnormal market event may result from many factors which are hidden in a large number of data. It is difficult to extract useful information from a large amount of sparse text [3]. Due to mentioned limitation, many existing machine learning methods are not suitable for abnormal market trends detection.

The main contributions of our work can be summarized as follows:

- We collect a labeled dataset for abnormal market trends and financial text with time stamps from the Oriental Wealth[1]. Most of our data sets are now publicly available at https://github.com/Dolphin02/Temporal-RBM/tree/master/data. Besides, we represent the financial text as a matrix based on semantics.
- We design an approach to detect the abnormal market trends. As an abnormal market trend event corresponds to too many texts so that we can not represent features efficiently, we first use a unsupervised Boltzmann method to embed many documents composed of the word segmentation into a feature vector and then use supervised machine learning methods to detect the abnormal market trends.
- We perform an experiment on the detection of abnormal market trends with financial texts. Compared to existing methods, our approach achieves state-of-the-art performance.

This paper is structured as follows: Introduction to study the background of this paper and methodology. In Section 2 we will introduce the related work in the fields of unusual fluctuating stock and text mining in the context of financial text. Furthermore, we will make a detailed description of the dataset we use in Section 3. Our methods for unusual fluctuating cause detection we proposed is given in Section 4. Then, Section 5 will introduce our experiment setup, compare the performance of our model with baseline models and presents our conclusions. Finally, we will conclude our research and propose our future plan in Section 6.

[1]http://www.eastmoney.com/

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

TABLE I
EXAMPLES FOR ABNORMAL TREND EVENTS

Stock ID	Abnormal trend type	Label	Date
300026	The value of the increase is greater than 7%	1	2017-11-24
600617	The value of the decline is greater than 7%	2	2014-12-17
600380	Within 3 trading days, a stock of 20% of the increase	3	2015-04-15
000386	Turnover rate is greater than 20%	4	2009-07-29

II. RELATED WORK

In this section we will explain the concept of abnormal market trends and financial text mining. Besides, we will formulate the correlations with financial text and abnormal market trends.

A. Abnormal market trends

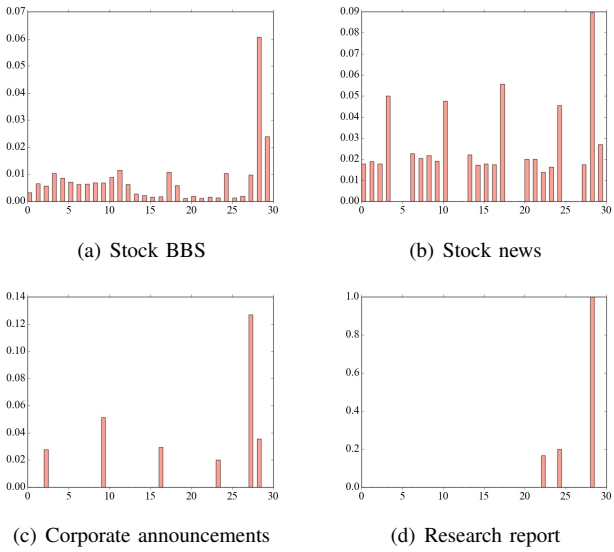
Nowadays, not a few stockbrokers attempt to model market trends by analyzing huge amounts of historical transaction data. For instance, some investment corporations run complex and time consuming simulations for the purpose of assess whether there will be abnormal trends. Katz et al. [4]count the frequency and magnitude of abnormal return events after Supreme Court action and suggests that there exists a "law on the market" effect which is measured by the frequency of abnormal return events.consider related approaches for tracking the impact of news on abnormal stock returns. There is no doubt that the emergence of market trends constitutes a potential risk and gain. Wang and Changyun [5] suggest that there exists close relationships between extreme sentiment and future market trends. Chaudhury et al. [6] find out that stocks tend to overreact for short periods subsequent to the publication of both positive and negative events that include global and domestic shocks.

B. Financial text mining

In this paper, Financial text data is composed of financial news, corporate disclosures, stock BBS (Bulletin Board System), public emotions or mood states and stock research report. Financial text contain rich information that having good link with several financial events or stock trend trends. Therefore, it is meaningful to extract some important features from financial text.

Not a few experts do some research on financial news. Schumaker et al. [7] investigate the relations between breaking financial news and stock price changes. Radinsky et al. [8] predict the stock price trend through stock news texts. In addition, Atkins et al. [9] suggest that using financial news predicts stock market volatility better than close price. The authors [10] analyze the effects of topics extracted in corporate disclosures on stock market returns and indicate that some topics have close relationships with returns. Yang et al. [11] study the relationship between news sentiment and market trends and suggest that the strategy generated based on the abnormal news sentiment methods has a good performance.

Not only do financial news affect the stock market, but also public emotions or mood states of stock BBS play an import part. Sentiment Analysis is widely used in text processing. Financial economics suggests that financial behavior and decision-making can be profoundly influenced by emotion and mood [12]. This is the reason why the public emotions or mood states can influence stock market trend as much as news. Gilbert et al. [13] extract the features of public anxiety from posts and investigates whether its changes can predict S&P500 values. Bollen et al. [2] indicate that combining public mood dimensions can improve the accuracy of DJIA predictions significantly.



(a) Stock BBS (b) Stock news

(c) Corporate announcements (d) Research report

Fig. 1. The changes in the posts of the stock 30 days before the abnormal fluctuations

III. DATA SET

A. Abnormal trend events

We build a crawler to collect abnormal trend events from winner list of Oriental Wealth. Each event is composed of trend type, stock id and the date of the trend occurred. The detail description of abnormal trend events is listed in Table I. We aligned the financial text with abnormal trend events according to when the trend occurred and when the articles were released.

TABLE II
THE DESCRIPTION OF FINANCIAL TEXT

Stock ID	Title	Content	Date	Clinic
60060	海信优势未来可期 *Hisense's advantages can be expected in the future*	海信电器公布了上半年财报，上半年销售收入微涨 2%……*Hisense Electric announced its financial statements for the first half of the year, Sales revenue increased slightly by 2% in the first half of the year*	2017-09-29	401
000816	周一开盘再跌停一次 *A limit-down on Monday open again*	大股东深陷资金危机，上周五跌停一次，抄底的不少……*The major shareholder was deeply involved in the financial crisis. Last Friday, there was a lot of bottom hunting.*	2018-01-26	1377

B. Financial text

In order to get the emotion information of the stocks, we obtained a collection of financial text composed of corporate announcements, stock news, stock research reports and stock BBS from Oriental Wealth, which was recorded from May, 2010 to January, 2018(3 million articles about 128 stocks). Each record provides an article identifier, the stock ID, the date of the release, its URL and the text content. We build a financial text crawler that extracts relevant market-related articles according to the stock ID and store them in the database. The crawler features a platform based on Python 3.6.2 that utilizes stock ID as a pre-specified keyword and records attributes such as the title, text content, release date, summary, click rate and the URL. The content of financial text are shown in Table II.

Due to the complexity of the textual structure, we first decompose the raw text into individual words with the Jieba[2] segmentation(the Chinese word segmentation module based on Python). Then we group all texts that were released on the same date and have same stock ID after the removal of stop words. Besides, we evaluate the frequency of financial text released daily and draw the changes in the posting amount of the stock 30 days before the abnormal fluctuations.

As shown in Fig. 1, we can find that users are more active before the abnormal trends occurred than usual. Since stock news, corporate announcements and research report are not released every day, the number of posts is 0 on some dates as shown in Fig. 1(b),(c),(d). Moreover, we can find some periodicity in Fig. 1(c). Because the stock does not trade on weekends, the activity of the investors is lower at every weekend.

C. Shanghai Composite Index

We obtained a collection of Shanghai Composite Index, defined to reflect daily changes in stock market value, over the same period of financial text release from SINA FINANCE[3]. Each data contains the date, opening price, closing price, ceiling price, floor price, the previous day's closing price, volume and (amount of increase and amount of decrease)Change. The k-line of the Shanghai Composite Index are shown in Fig. 2.

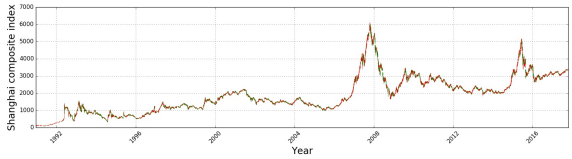


Fig. 2. The volatility of Shanghai Composite Index

D. Data preprocessing

At first, we train the word2vec according to the parameters settings in Table III. The available word vectors are trained on our financial text dataset and are of length 128. Then we represent financial text composed of words with sentiment lexicons and stored them in a document.

IV. METHOD

In this section, we first describe the problem we studied and formulate our model, followed by the text data representation and the time series feature extraction module. Finally, we describe our method to classify stock trends.

A. Problem description

Motivated by the previous findings of stock trading events, the majority of meaningful abnormal market trends can be observed within a short period after the release of correlative information [14] and the correlation weakens as time goes [15]. The goal of our study is to investigate whether financial articles could bring about significant impact to abnormal market trends. We are interested in identifying abnormal sentiments that may lead to abnormal market trends and proving our model is sound by evaluating the performance of abnormal trend classifications rather than proposing an optimal abnormal trend prediction model.

We use a data set that is composed of financial text(denoted by X), the amount of posting(denotes N, where $N = \{n_1, n_2, \ldots, n_{10}\}$) and Shanghai composite index((denotes P, where $P = \{p_1, p_2, \ldots, p_{10}\}$)) within 10 days before the occurrence of the trend as input. Financial text X contains X_c(Corporate Announcements), X_r(Research Report), X_n(Stock News) and X_b(Stock BBS). We build a model that can classify the type of the abnormal market trend.

We considered four abnormal market trends. The meaning of its values is listed in Table I. Our target is building a model to detect whether there exists an abnormal market trend and the type of abnormal market trend if it occurs at time t using the historical values at time $\{t - 1, t - 2, t - 3 \ldots t - 10\}$. In order to test whether our model is effective, we compare our method to extract textual sentiment with baseline models.

[2]https://github.com/fxsjy/jieba
[3]http://finance.sina.com.cn/stock/

TABLE III
MAIN NOTATIONS DEFINITIONS AND SETTINGS IN WORD2VEC

Notion	Definition	Values
cbow	Represents whether to use cbow model, 1 means to use cbow model, 0 means to use skip-gram model	0
size	Represents the output word vector dimension	128
window	Represents the size of the training window	5
negative	The number of words contained in the generated negative sample NEG(w)	25
hs	Represents whether to use hs model, 0 means not use	0
binary	Represents whether the output file is binary storage, 0 means not use, 1 means use	1
sample	Represents the sampling threshold	1e-4

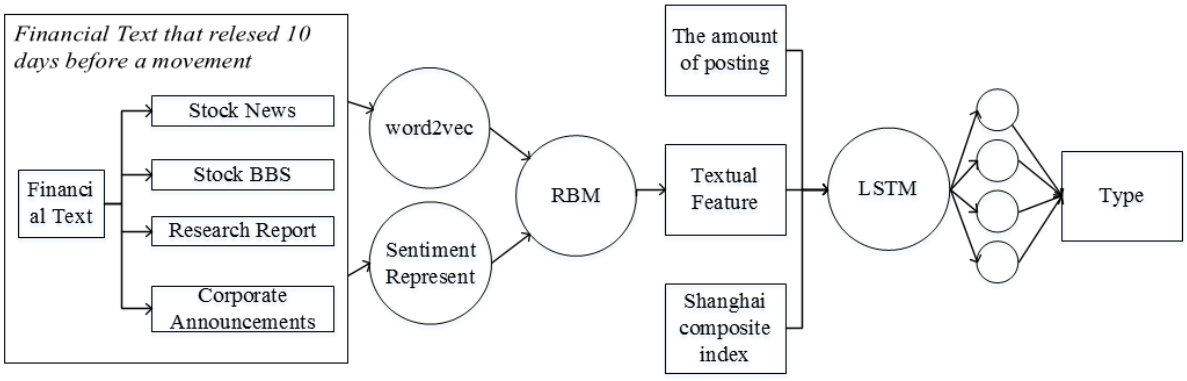
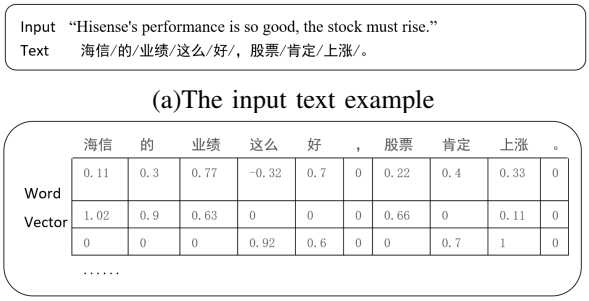


Fig. 3. The framework of our approach



(a)The input text example

(b)The text composed of word vectors

Fig. 4. Word Embedding

B. Proposed approach

As shown in Fig. 3, we design a general framework which is suitable for observing abnormal market trends with financial texts and other time series data. We first convert the financial texts X to structured word vectors W and then extract the emotional features, denoted as S, from four types of financial text, then we combine the changes in the number of postings and the features of the market index with S and input them together into LSTM (Long Short-Term Memory). Then input the outputs of LSTM to a softmax layer. Finally, we can get the classification results of abnormal trends.

1) Word embedding: Since current effective textual mining algorithms are not able to cope with plain text as input, each word need to be performed some preprocessing to transform textual documents into a feature vextor. Traditional lexicon-based methods exist that some words are missing in current lexicons. And n-grams treat words as discrete elements, which would result in a high dimensional vector and do not take the continuity of semantics into account. In this paper, we mainly use the popular approach [16] in the text representation and use the sentiment lexicons method as the additional sentiment dimension. The description of these two methods are as follows:

- **Word2vec** The word2vec tool is fast and currently widely used in measuring syntactic and semantic word similarities. We use Skip-Gram model in our study. In Skip-Gram model, the Huffman code for each word is used as an input to a log-linear classifier with a successive projection layer and predicts words within a given context window [17]. The available vectors are trained on our financial text dataset, composed of 64 million words, and are of length 128[4]. The training parameters are shown in

[4]https://code.google.com/p/word2vec

TABLE IV
THE RULES OF EMOTIONAL FEATURE CONSTRUCTION

Meaning	Value	Meaning	Value
positive emotional word	0/1	negative emotional word	0/1
negative word	0/1	adverb of degree	0/1
noun	0/1	verb	0/1
adjective	0/1	adverb	0/1
Is it emotional punctuation	0/1		

Table III.

- **Representing with sentiment lexicons** In order to extract more features from financial text, we constructing emotional dictionaries using How-net emotional word set[5] and design several construction, illustrated in Table IV. Based on the words after the word segmentation, query whether these words will appear in the positive emotion dictionary, the negative emotion dictionary, adverb of degree dictionary and the negative dictionary. If they appear, the corresponding position is assigned a value of 1. Then we assign values to nouns, verbs, adjectives and adverbs according to the part-of-speech and construct each sentence as an emotional feature matrix.

After converting words into vectors, we can get a structured vector sets where we use the semantics and emotions of words to represent the words themselves. In this step, we convert the financial texts in Fig. 4(a) to structured sets W in Fig. 4(b). Then we can use machine learning methods to operate on these texts composed of word vectors.

2) Document embedding: The original documents composed of word vector matrixs are so sparse that a large number of parameters have to be leraned. Therefore, the current machine learning algorithms and normal equipments cannot calculate at a fast speed. Therefore, we need to reduce the dimension of the data by embedding word vector to document vector.

We initially use the Doc2Vec method proposed by Quoc Le and Tomas Mikolov [18] that already integrated in gensim[6], but the performance on our dataset is not good. RBM (Restricted Boltzmann Machine) is an unsupervised machine learning model, which is used to reconstruct the input data, that is, to effectively extract data features, construct a new data structure for predictive analysis. So we embed financial texts that were released on the same date into a vector using the unsupervised RBM, which perform well.

The RBM layer in our model is similar to that in the deep belief net (DBN) proposed by Hinton et al [19]. As shown in Fig. 5, we input a financial texts represented with above mentioned word vector matrixs W to the input layer, then calculate hidden layer values. We are able to obtain a document vector D contains important text features after training many times.

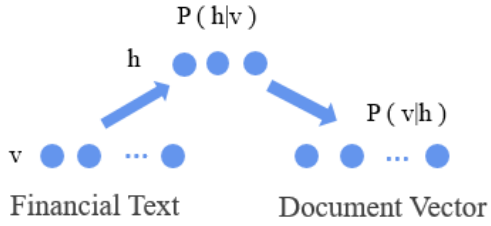


Fig. 5. Doc2Vec architecture

3) Trend Type Classification Model: LSTM is very popular in processing text data, which has achieved pretty significant results in sentiment analysis, machine translation and text generation. Because the problem involves similarly classified sequences, LSTM is an excellent method in this situation.

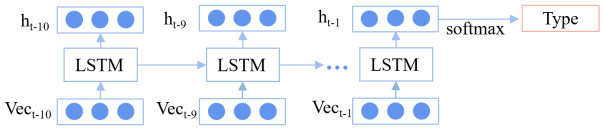


Fig. 6. Trend type classification architecture

We add the daily posting amount and Shanghai composite index over the same period as the additional feature dimensions to the document vectors D and mark the new mixed vector as m_X. As shown in Fig. 6, then we input the mixed vector m_X that within 10 days before an abnormal trend occurs into LSTM. At last, we input the outputs of LSTM to a single-layer neural network to determine the type of abnormal trends.

V. EXPERIMENT

In this section, we first introduce our data set division and evaluation of the results, then we describe the process of parameter selections and finally explain our performance to extract the features of the text. The complete source code and data sets are available at https://github.com/Dolphin02/Temporal-RBM.

[5]http://www.keenage.com/html/c_bulletin_2007.htm
[6]https://radimrehurek.com/gensim/models/doc2vec.html

TABLE V
THE RESULTS(%) OF THE TRENDS CLASSIFICATION

Type	Baseline Models												Our Model			
	SVM				NB				Random Forest							
	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1
1	72.1	72.3	83.4	77.5	69.3	71.5	79.9	75.4	73.2	74.4	88.1	80.6	86.1	87.3	95.6	**91.2**
2	64.1	65.0	80.3	71.8	69.2	69.7	80.9	74.8	63.9	64.5	88.2	74.5	75.6	75.2	**96.7**	84.6
3	61.1	42.3	39.8	41.0	54.4	48.3	54.2	51.1	59.6	52.3	48.7	50.4	84.6	82.9	74.5	78.5
4	54.1	54.9	35.5	43.1	57.2	53.3	57.9	55.5	57.8	58.6	29.3	39.1	**88.3**	**90.2**	85.1	87.6

A. Experiment setup

We divided samples into two parts randomly: two-thirds for training, and one third for testing. We assigned each sample a label by comparing its stock code and date with the winners list, give them different labels according to different reasons for the winners list, such as the stock rose more than 7%. We evaluate the performance by using the accuracy metric.

1) Evaluation: In this paper, we use the accuracy, recall, precision and F1-measure four metrics to measure the performance of the trend type classification. Given the collections of N samples, which contains the predicted result is positive, and it is actually positive (TP), the predicted result is negative, and it is actually negative (TN), the predicted result is positive, but it is actually negative (FN) and the predicted result is negative, but it is actually positive (FP).

TABLE VI
CONFUSION MATRIX

	Actual "positive"	Actual "negative"
Pre "positive"	TP	FN
Pre "negative"	FP	TN

The confusion matrix table shown in Table VI is used to calculate the mentioned performance measures. Equation (1) is used to calculate the accuracy. The accuracy measures whether the classification is correct. Equation (2) is used to calculate the precission, which shows how many perdicted positive samples are truly positive samples. Equation (3) is used to calculate the recall, which indicates how many examples in the original positive samples were correctly predicted. As there is a trade-off between recall and precision, the figures shall not be assessed in isolation. Thus the F1-measure additionally in (4). The higher the metrics, the better the classification effect.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

2) Baseline: In order to access the effectiveness of our model, we compare the classification accuracy on three kinds of baseline model:

- **SVM** (Support Vector Machine) model: SVM, proposed by Vapnik [20], is considered as one of the most robust and accurate methods among the well-known data mining algorithms [21]. Due to its effective performance in solving non-linear problems, SVM is widely used in research and industry [22].
- **RF** (Random Forest) model: The random forest classifier, proposed by Breiman and Leo [23], consists of a combination of tree classifiers. Each classifier is generated by sampling independently from the input vector using a random vector and each tree casts a unit vote for the most popular class to classify an input vector.
- **NB** (Naive Bayes) model: McCallum et al. compare different models for naive Bayes on text classification and indicate that the multinomial performs well [24], so we choose it as our baseline model.

B. Experiment results

This experiment was performed on the tensorflow 1.4.1 framework and run in Ubuntu 16.04.1 system. In document vector embedding, the number of visual layer nodes is the same as the dimension of the original vector formed by the word vectors and the number of hidden nodes is 100. Besides, we set the learning rate to 0.01. In the abnormal emotion feature extraction section, we adopt AdamGrad to train our model and set the learning rate to 0.01. During training, we adopt the dropout operation, the probability is 0.5, before the softmax layer.

The performances of the trend type classification are shown in Table V, which our model has obviously achieved state-of-the-art performance on our Chinese financial text dataset described in Section 3. This experiment confirm that when it is necessary to perform feature extraction on a large amount of sparse text data, as the machine learning algorithms cannot be used, we can add a RBM layer in front of the machine learning model to reduce the dimension. In addition, the results show that our model are novel and effective for detecting the abnormal market trends using abnormal sentiment information

extracted from financial text. Therefore, we can conclude that there exists a strong correlation between the release of the financial text and abnormal stock market trends.

VI. Conclusion and future work

Abnormal market trends detection of Chinese stock market is a challenging task, because the stock trends are affected by many factors and the lack of labeled financial text that can be trained. The main contribution of this study can be summarized as follows. Firstly, while the lack of labeled financial text, we build some crawls to obtain financial text dataset to train. Besides, we proposed a novel hybrid model to capture the abnormal sentiment in social media for the detection of abnormal market trends on a large amount of financial text. Finally, we reported the effectiveness of detecting abnormal market trends on our dataset by performing experiments on a large scale test data. A limitation of our study is that we are not able to detect some trends that occurred less frequent. To overcome this weakness, a bigger more comprehensive data sets will be collected and trained to extract abnormal sentiment for the abnormal trends detection.

For future work, we will try to integrate more factors that can lead to the abnormal stock market trends to develop a more accurate trends detection model. Besides, our current study can be extended to investigate the correlations between financial text and market risk management because we foresee that the market risk may take the market trends in this study into consideration. In addition, we can investigate abnormal market trends to formulate diversified trading strategy in the future.

Acknowledgment

This work is supported by the National Key Research and Development Program of China under Grant No.2017YFB1401202 and the Key Research and Development Program of Zhejiang Province under Grant No.2017C01013.

References

[1] C. Kenny, "The internet and economic growth in less-developed countries: A case of managing expectations," *Oxford Development Studies*, vol. 31, no. 1, pp. 99–113, 2003.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[3] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[4] D. Katz, I. Bommarito, J. Michael, T. Soellinger, and J. Chen, "Law on the market abnormal stock returns and supreme court decision-making," *ArXiv Preprint V2*, 2015.

[5] C. Wang, "Futures trading activity and predictable foreign exchange market movements," *Journal of Banking & Finance*, vol. 28, no. 5, pp. 1023–1041, 2004.

[6] M. Chaudhury and P. Piccoli, "How do stocks react to extreme market events evidence from brazil," *Research in International Business and Finance*, vol. 42, no. 2017, pp. 275–284, 2015.

[7] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system," *ACM Transactions on Information Systems*, vol. 27, no. 2, p. 12, 2009.

[8] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 909–918.

[9] A. Atkins, M. Niranjan, and E. Gerding, "Financial news predicts stock market volatility better than close price," *The Journal of Finance and Data Science*, 2018.

[10] S. Feuerriegel, A. Ratku, and D. Neumann, "Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation," in *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, 2016, pp. 1072–1081.

[11] S. Y. Yang, Q. Song, S. Y. K. Mo, K. Datta, and A. Deane, "The impact of abnormal news sentiment on financial markets," *Journal of Business and Economics*, vol. 6, no. 10, pp. 1682–1694, 2015.

[12] J. R. Nofsinger, "Social mood and financial economics," *The Journal of Behavioral Finance*, vol. 6, no. 3, pp. 144–160, 2005.

[13] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," in *ICWSM*, 2010, pp. 59–65.

[14] B. F. Smith, R. White, M. Robinson, and R. Nason, "Intraday volatility and trading volume after takeover announcements," *Journal of Banking & Finance*, vol. 21, no. 3, pp. 337–368, 1997.

[15] J. M. Patell and M. A. Wolfson, "The intraday speed of adjustment of stock prices to earnings and dividend announcements," *Journal of Financial Economics*, vol. 13, no. 2, pp. 223–252, 1984.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ArXiv Preprint ArXiv:1301.3781*, 2013.

[18] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[19] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[20] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[21] X. Wu, V. Kumar, and J. R. Quinlan, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[22] H. X. Zhao and F. Magouls, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI Workshop on Learning for Text Categorization*, vol. 752, no. 1, 1998, pp. 41–48.