Analysis of Relationship Between Hot News and Stock Market——Based on LDA Model and Event Study

Chanyuan Liu
School of Economics And Management, Beijing Jiaotong University, Beijing 100044, China
Acelinliu0908@163.com

Abstract. This paper clustered the document of news released by major mainstream media outlets in November 2019 through LDA topic model and finds out the hot topic this month through the probability distributions between the topics and within the words. Then we backtracked on the topic and find the initial occurrence of the news, then categorized the news according to the status of the facts reported in the news. Then we carried on the research through using the constituent stock of concept stock in this model to report the event in a short-term event study, which finally find the breaking news could cause the related module in the short-term abnormal return while the continuing news would not do. It means that breaking news could allow media report to keep pace with market reactions, which the continuing news cannot promised. Investors will not be affected by continuing media reports on investment judgment as they have received relevant information in advance.

1. Introduction
Development of dissemination of news has been a long process from the beginning of the embryonic state to the formation. The word “new” in news has plenty of explains, which can be seen as the new of timeliness or unconventional content. However, no matter how peculiar and novelty it is, it is based on the fact, which could not be fabricated. Therefore, the effect of dissemination of news should be analyzed based on the status of facts of the news. [1]

In recent years, the researches of the combination of media effect and stock market are increasing. The earlies paper of literature in this category included Klibanoff [4], who mentioned the specific country/region news reported on the front page of The New York Times has influenceed the pricing of closed-end country/region funds. They found that price movements are closely related to fundamentals within weeks of the front-page news. Tetlock [5] analyzed the language content of mass media and reported that the pessimism of media predicted the down stress and the subsequent changeover.

As the aspect of empirical study, Fand and Peress [3] researched the relationship between media attention and the return of stock intersecting surface based on the securities market in America, which showed that media attention and stock returns are inversely variable. However, Bushee et.al [6] showed that intensive media coverage can reduce the degree of information asymmetry, which can increase stock market liquidity and stock market trading. The share price and the volume of business would increase of the enterprise, which had been closely covered by the media.

With the development of the modern science and technology, text mining technology, machine learning and natural language processing methods are also constantly improving, which makes it easier to process huge amounts of information in a big data environment. For example, machine learning can categorize and classify news texts in terms of analyzing news emotions, then the
classification can be divided into positive emotion, negative emotion and neutral emotion if KNN and SVM arithmetic are used, and it can also build an emotional index from each quantity of category. YouJiaxing and WuJing[14] found that the higher the negative or positive emotions in news reports, the more likely the stock price is to deviate from its original value by using quarterly data from listed companies.

LDA topic model has attracted more and more attention in the field of natural language processing in recent years. Blei [10] and so on proposed the LDA (Latent Dirichlet Allocation) model in 2003, which treats a topic as the probability distribution of words. After a document is trained by the LDA model, its representation in the topic space can be obtained, and eventually generate each topic and the probability distribution of words of different terms in each topic [16].

This paper has the following contents in the research based on the review of literature above:
(1)Clustering text of a large number of daily news by using the LDA topic model to obtain the hot news issues within a period.
(2)Financial news is divided into breaking news and continuing news according to relevant theories of journalism and communication. And discussing the co-movement relationship between hot news and stock market through researching the classification of the short-term events of different news.

2. Research design and research hypothesis
The basic thoughts of this paper are build LDA probabilistic thematic model through python to cluster the hot news in November, 2019 (before and after 5 trading days). It divides hot news into breaking news and continuing news, analyses the events of different kinds, and discusses the short-term market reaction of relevant sections of that hot news.

2.1. Research Method

2.1.1. LDA topic model
LDA(Latent Dirichlet Allocation): The potential Dirichlet distribution is an essential unsupervised algorithm in the field of NLP (natural language processing), which basic thought is the inverse of the text generation process. It means that searching the theme as well as the words corresponding of these D existing documents. Blei [10] proposed the LDA model in 2003, which expressed the probability distribution of the document theme by using K - dimensional implicit random variables obeying Dirichlet distribution, based on pLSI. After this, Griffiths [11] threw Dirichlet distribution on β parameter, which made LDA model become a complete one. (See Figure 1):
When determining the number of topics $K$, this paper chooses to determine the most appropriate $K$ value according to the way of calculating the perplexity from Blei[10]. The calculation formula of perplexity is as follows:

$$\text{perplexity}(D) = \exp\left(\frac{\sum D \log p(w)}{\sum D N_d}\right)$$

(1)

Using text description to describe the LDA model could be understood as the degree of uncertainty about those topics in document No. $D$ in model training. The lower the perplexity is, the better the clustering effect of the model.

### 2.1.2. Event study method

The process of the event approach can be roughly summarized as: first determining the window period and estimating the expected return rate of it through the window period samples. Then, subtracting the normal return to obtain abnormal return (AR) in the effective return rate in window period. Finally, examining if the cumulative average abnormal return (CAAR) over the sample window period is significantly different from the null hypothesis (cumulative abnormal is equal to zero).

This paper selected hot topic through LDA model and look back upon the time of it happened, finally determined the window phase was $(-150, -30)$. After using the related plate to get rid of the residue constituent stock of ST stock, we obtained the average abnormal return through differentiating the cross section of average of abnormal return within $[-1,1]$ window period to explore the market's short-term reaction to different news events.

### 2.2. Theoretical basis and research hypothesis

As far as this paper concerned, the media report more times and higher frequency, which could create a stronger topic in the social public opinion. Investors are more likely to react in the market if the news is repeated many times. So, based on this, clustering analysis the media report through LDA model can react the heat of a certain topic.

According to "Introduction to Journalism" by LiLiangrong[2], this paper classified news into breaking news and continuing news, the breaking news is report for the issues, which is out of expectation and breaking out suddenly. And the continuing news is reported for the issues that are gradually changed. For a certain topic, the initial report should be the breaking news, and the subsequent report should be the continuing news. Breaking news is unpredictability, hence the media report is firsthand information, which will be reflected in short-term market abnormal return after received by investors. But the relevant mode will not react for continuing news. So, this paper mentions hypothesis below:

**Hypothesis 1**: Breaking news will produce remarkable abnormal returns in the short term in hot topic.

**Hypothesis 2**: Continuing news will not produce remarkable abnormal returns in the short term in hot topic.

### 3. The empirical results

#### 3.1. The result of LDA topic model

This paper used 52,699 news reported by the major mainstream media from CSMAR database from 15:00 October 31, 2019 to 15:00 November 30, 2019. Using the python3.7 and program the LDA model to extract the topic information. The number of topics $K$ is a very important parameter when using the LDA model. We use the perplexity to determine the topics number $K$. We could see that in Figure 2, when $K=9$, the perplexity is the lowest. Therefore, this paper uses $K=9$ as the number of topics in the hot news of the month.
After obtain the best number $K=9$ by LDA model, this paper gets 9 top news topics of the month in total. The relevant information is shown in Table 1.

### Tab.1 topics-key words and proportion of topics

| Topic number | Keyword1    | Keyword2   | Keyword3 | Keyword4      | Theme proportion |
|--------------|-------------|------------|----------|---------------|------------------|
| 1            | cooperation | China      | strategy | A-stock       | 0.1716           |
| 2            | High-speed  | business   | IPO      | listed        | 0.1341           |
| 3            | 5G          | generation | construction | launch    | 0.1166           |
| 4            | trillion     | level      | electronic | GDP        | 0.1090           |
| 5            | supervise    | environment | rating | hangtag | 0.1098           |
| 6            | expect       | futures    | pig      | industrial | 0.1004           |
| 7            | China        | central bank | Repurchase | interest rate | 0.0969           |
| 8            | interim      | Hongkong   | IPO      | Alibaba     | 0.0930           |
| 9            | innovate     | convention | mechanism | system     | 0.0764           |

Note: In this table, the occurrence probability of key words is displayed from high to low according to the proportion of subject in the results, and keywords similar to theme judgment are excluded, such as "company", "fund" and "stock price".

According to the consequence, we looked back upon the original document of the news to find out the representative news and record the time of press release. This paper eliminated some ambiguous topics, selected more representative topics 2, 3 and 8 after screening and classified relevant news on this topic as shown in Table 2.

### 3.2. Event study results and robustness tests

This paper classified news into breaking news and continuing news, and used the event study method to respectively select the information of constituent stocks of subway mode from sequence (except ST), 5G concept stock mode (except ST) and Alibaba concept stock from CSMAR, which study the short-term issues of those three topics. And it calculating the average excess return CAAR of concept stocks in the window period by using Stata software.

### Tab.2 Backtracking the specific news and time of different topics

| Topic number | Keyword | Variable name | Related News | Occurrence time |
|--------------|---------|---------------|--------------|-----------------|

...
This paper uses market model to estimate the normal return. And it was selected by composite A stock return from the market. In this paper, the selected window period is (-150, -30) and the event window period is [-1,1], the significance test is conducted on the results. The empirical results are shown in Table 3:

It can be seen from Table 3, the abnormal short-term market returns caused by breaking news are significant, which indicates that breaking news is more likely to have an impact on the market. However, the continuing news did not have within the scope of investors' expectations, whether the information of media report is official or not, it cannot get extraordinary feedback in the market. For example, fits abnormal earnings were not significant when Alibaba went public, which reaction on this day was far less than the media's reaction on November 8, when it was rumored that the "road show" was about to begin. Even some variables reached the opposite conclusion. (like Conti4 and Conti7). This also verifies the hypothesis proposed at the beginning of this paper.

| variable | occurrence time | CAAR[-1,1] | CAAR[-2,1] |
|----------|-----------------|------------|------------|
| Break1   | 2019-11-07      | -0.006673** | -0.0053949* |
|          |                 | (0.049)    | (0.098)    |
| Conti1   | 2019-11-11      | -0.0029285 | -0.002408  |
|          |                 | (0.445)    | (0.643)    |
| Conti2   | 2019-11-14      | 0.00228    | 0.004301   |
|          |                 | (0.630)    | (0.412)    |
| Break2   | 2019-10-25      | 0.0046074* | 0.0016738  |
|          |                 | (0.094)    | (0.179)    |
| Conti3   | 2019-11-04      | -0.0003263 | 0.0044851  |

Tab.3 Results of event study
4. Conclusion

This paper clusters the document of news released by major mainstream media outlets in November 2019 through LDA topic probability model and finds out the hot topic of that month. And it picks up the model topic to recall the specific occurrence time of the original news and classifies the news through the fact status and uses the constituent stock of the relevant concept module to research the abnormal return of the short-term within [-1,1] for that reported news. However, the continuing news hasn’t enough evidence to prove that it can causes abnormal returns in related plate in the short term.

It means that the media reports timely and the reaction of market can also be explained by the news contents. However, as for continuing issues, the timeliness of media reports is not high, which indicating that investors have received relevant information before reporting. Share price information already includes those from media reports. Therefore, the media report at this time cannot influence the investment judgment of investors. This is also the certain research significance for the selection of investors for media reports.

Acknowledgments

This paper was supported by Qi He, thanks her for the knowledge of journalism. Also, I would like to thank Lily Zhang for her helpful comments and suggestions.

References

[1] Chen Lin. Introduction to Journalism [M]. Suzhou: Suzhou University Press.2007:14-16. (Chinese Version)

[2] Li Liangrong. Introduction to Journalism [M]. Fudan University Press: Shanghai,2001:32-36. (Chinese Version)

[3] LILY FANG,JOEL PERESS. Media Coverage and the Cross-section of Stock Returns[J]. The Journal of Finance,2009,64(5).

[4] Peter Klibanoff, Owen Lamont, Thierry A. Wizman. Investor Reaction to Salient News in Closed-End Country Funds. 1998, 53(2):673-699.

[5] PAUL C. TETLOCK. Giving Content to Investor Sentiment: The Role of Media in the Stock Market[J]. The Journal of Finance,2007,62(3).

[6] Brian J. Bushee, John E. Core, Wayne Guay, et al. The Role of the Business Press as an Information Intermediary. 2010, 48(1):1-19.

[7] JOSEPH E. ENGELBERG,CHRISTOPHER A. PARSONS. The Causal Impact of Media in Financial Markets[J]. The Journal of Finance,2011,66(1):.

[8] Janusz Brzeszczyński, Jerzy Gajdka, Ali M. Kutan. Investor response to public news, sentiment
and institutional trading in emerging markets: A review[J]. International Review of Economics and Finance, 2015, 40.

[9] Paresh Kumar Narayan, Deepa Bannigidadmath. Does Financial News Predict Stock Returns? New Evidence from Islamic and Non-Islamic Stocks[J]. Pacific-Basin Finance Journal, 2015.

[10] Blei D, Ng A, Jordan M. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003: 993-1022

[11] Griffiths Thomas L, Steyvers Mark. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 Suppl 1.

[12] Huangjun, GuoZhaorui. News Media Report and Capital Market Pricing Efficiency -- Analysis based on Stock Price Synchronization[J]. Management World, 2014(05): 121-130.

[13] LuoJinhui, DUXingqiang. Media Report, Institutional Environment and Stock Price Crash Risk[J]. Accounting Research, 2014(09): 53-59+97. (Chinese Version)

[14] YouJiaxing, Wujing. Spiral of Silence: Media Sentiment and Asset Mispricing [J]. Economic Research, 2012, 47(07): 141-152. (Chinese Version)

[15] WangChangyun, WuJiawei. Media Tone, Investor Sentiment and IPO Pricing [J]. Financial Research, 2015(09): 174-189. (Chinese Version)

[16] Xuge, WangHoufeng. Development of Thematic Models in Natural Language Processing [J]. Journal of Computer Science, 2011, 34(08): 1423-1436. (Chinese Version)

[17] YuanXianping, KeDagang. Event Research Method and Its Application in Financial Economic Research [J]. Statistical Research, 2006(10): 31-35. (Chinese Version)