Content analysis and topic correlation of financial news related to big data concept

Yuanyang Teng, Xiaobo Wu, Yicun Li
School of Management, Zhejiang University, Hangzhou 310058, China
wuxb@zju.edu.cn

Abstract. This paper combines the traditional content analysis method with the emerging natural language processing algorithm, LDA, and takes the financial news related to big data as the research object of content analysis, so as to extract the topics and vocabulary concerned by different categories of financial news related to big data, and judge the correlation of different categories of financial news on topics.

1. Introduction

Traditional content analysis methods convert non-quantitative literature materials into quantitative data, make quantitative analysis of literature contents and judge as well as inference about facts based on these data. Moreover, its analysis of the factors and structures composing the literature is more detailed and procedural. The general process of content analysis includes the establishment of research objectives, determination of research population and selection of analysis units, design of analysis dimension system, sampling analysis process and quantitative analysis of materials, evaluation record and analysis inference. These six parts have been programmed and quantified to a certain extent, but the quantified procedures are mainly at the dimension perspective and sample perspective, and have not gone deep into the basic unit of text -- word level. Based on the general steps of content analysis, this paper first refers to content analysis method on the dimension design of big data-related news categories, and then divides the content of each news item into words, finally, LDA algorithm is used to carry out topic modeling at the word level, and relevant conclusions are drawn. This kind of in-depth study of massive text is beyond the traditional content analysis method and is also the innovation point of this paper.

2. Related Literature

In the way of information transmission in contemporary society, the network has become an important data source for all kinds of analysis. People use network data for public opinion analysis, sentiment analysis and so on. Li Shi et al. [1] analyzed the characteristics of the product and the public's emotional tendency towards the product by using people's comments on the Internet. Some scholars have begun to pay attention to the analysis and prediction of the stock market by using network data. Johan B et al. [2] extracted emotions from various states posted by people on Twitter and made corresponding statistics, and made effective predictions for the stock market, which have been practically applied to a quantitative fund abroad.

From the original “Latent Semantic Indexing” to “Probabilistic Latent Semantic Indexing”, to the “Latent Dirichlet Allocation”, the semantic mining of topic model on text has become more mature. Wei X et al. [3] used LDA to analyse queries and texts in information retrieval and combined it with
vector space model to obtain a new model. The performance of this model in information retrieval is higher than that based on clustering. Zhang Jianwei et al. [4] modified the LDA model based on dynamic vocabulary, making it perform better in the confusion degree and mutual information index. Hao Jie et al. [5] gave different weights to different words in Gibbs sampling, which enhanced the influence of words with emotional tendency in the sampling process and thus improved the distinction between topics.

Zhang Shujun [6] started from the event of "Everbright Bank's black index", selected the financial topics of 21st century and NetEase as the research objects, and carried out comparative analysis from the layout design, content source and integration, interaction, multimedia application and other indicators. Through qualitative and quantitative analysis of two financial news articles in China Daily.

In this paper, by learning the above-mentioned literature of network information analysis, LDA related improvement model and financial news content analysis, combined with content analysis method and LDA, the financial news relating the concept of big data in China was studied as the object of study, and the results of topic correlation were analysed.

3. Data description
In this paper, 1025 pieces of big data-related financial news were selected from Chinese portals such as Baidu, 360 and Toutiao. For each piece of news, three fields, namely news headline, news source (or author) and news content, were recorded respectively. Among them, news content is the main object of this study, as shown in Figure 1.

We took out the 1025 pieces of news and split news into single words, and got the content as shown in Figure 2. These news words are the basic objects of follow-up research.

4. Research methods

4.1. Content Analysis
Content analysis is a research method which mainly takes various literatures as the research object. The early content analysis method originated from the quantitative analysis of historical documents by means of natural science research. Content analysis converts non-quantitative literature materials into quantitative data, and makes quantitative analysis, judgment and inference about facts based on these data. Moreover, its analysis of the factors and structures composing the literature is more detailed and procedural. The general process of content analysis includes six parts: establishing the research objective, determining the research population and selecting the analysis unit, designing the analysis dimension system, sampling analysis process and quantitative analysis material, evaluation record and analyzing inference.

Figure 1: Data description
Figure 2: News content and its split words
4.2. LDA (Latent Dirichlet Allocation)
LDA (Latent Dirichlet Allocation) is a document generation model. It believes that a text has multiple topics, and each topic corresponds to different words. In the construction of an article, first select a topic with a certain probability, and then select a word under this topic with a certain probability, thus generating the first word of this article. This process is repeated over and over again, and the entire article is generated. It is assumed, of course, that there is no order among words. The algorithm diagram of LDA is shown in Figure 3.

LDA is to add Bayesian framework on the basis of pLSA, that is, LDA is the Bayesian version of pLSA, and their framework steps to generate documents are the same. The difference is that pLSA, holding a frequency opinion, believes that topic distribution and word distribution are uniquely determined. However, as a Bayesian viewpoint, LDA believes that the distribution of topics and words is no longer uniquely determined and unchanged, that is, it cannot be given exactly. It is known that each article has its own topic, and the topic distribution is multinomial distribution, whose parameters follow Dirichlet distribution. Each topic has its own word distribution, which is also a polynomial distribution whose parameters follow the Dirichlet distribution. Simply put, LDA is adding two Dirichlet priors to the topic distribution and word distribution based on pLSA. The process of LDA is shown in Figure 4, which continues five process: selecting a document according to prior probability; sampling topic distributions of generated document are from the Dirichlet distribution A; generating a word of the document by sampling from the polynomial distribution of the topic; generating the word distribution corresponding to the topic by sampling from Dirichlet distribution B; generating the words by sampling from the polynomial distribution of the words.

5. Empirical research

5.1. Content Analysis
Based on traditional content analysis, this study classifies 1025 news articles into three categories: business, technology and society. The specific method is to randomly select the news headlines, classify the news into three categories based on the understanding of the headlines, and get the following table 1.

Through the analysis of table above, we found that, although we collected financial news under the big data keyword, but the main focus of social public opinion is not a simple (only 37%) accounted for technology, more is thinking about big data, artificial intelligence and other emerging technology impact on people's social life, social category accounted for the highest, up to 40.7%, and while news of finance and economics are collected, but big data related to business news proportion is the least (22%), big data and the commercialization of the artificial intelligence technology still has some way to fall to the ground.

| Table 1 Proportion of news categories in content analysis |
|----------------------------------------------------------|
| category | Headers sampling from 1025 news | Proportion in category |
|----------|--------------------------------|-----------------------|
|          |                                 |                       |
| Technology | 1. Overview of the latest big data concept stocks in the big data industry |
|------------|-------------------------------------------------------------------|
|            | 2. Google’s TensorFlow playground brings you neural networks!    |
|            | 3. Li Keqiang: China’s big data and cloud computing industries are open |
|            | 4. Three neglected points of ARTIFICIAL intelligence: subversion, self-evolution and de-nodal |
|            | 5. How to implement Serverless architecture in Aliyun plus platform? |
|            | 6. Big data promotes the construction of land engineering discipline |
|            | 7. Research on artificial intelligence and big data innovation-- Tsinghua big data "application innovation” lecture series | CWCISA recommendation |
|            | 8. SAP launches next generation data warehouse application SAP BW/4HANA |
|            | 9. Wang Chuan: How deep is deep learning? (24) Jobs and Deep mind break |
|            | 10. Big data and small data: what kind of method can solve what kind of problem | up to now see the most professional article |
| Business   | 11. Baidu Marketing Research Institute: How to make institutional marketing strategies by referring to the retrieval data of consumers on the whole web |
|            | 12. CIO: The difficulty of big data for business decisions |
|            | 13. Analysis of the latest product promotion data of One Mob advertising platform |
|            | 14. "Big data" has helped transform the world economy |
|            | 15. Tiki Data, an independent subsidiary of Cloud Intelligence, was jointly invested by Sequoia Capital and Gobi Venture Capital |
|            | 16. Internet advertising: The appearance level of big data realization |
| Society    | 17. What is the educational significance of teenage robots without exams? |
|            | 18. Tianrunhuabang + Provisions on Electronic Data Collection, Extraction and Judgment issued by supreme Law, Supreme People’s Procuratorate and Ministry of Public Security (effective October 1, 2016) |
|            | 19. Crossing the data tipping point for real-time insight into cognitive medicine |
|            | 20. Famous teacher talks about teaching | to teach students in-depth Chinese learning |
|            | 21. Read in a Minute The dilemma of deep learning overtaking human intelligence |
|            | 22. The incidence of myopia among Chinese college students is 90%. Chengdu will collect and analyze big data of "glasses |
|            | 23. An Tai ‘Mid-Autumn Tourism big data released, come to watch’ |

| 37%        | 22% |

| Business   | 22% |

| 40.7%      | 40.7% |
24. Typhoon "hit" Mid-Autumn Festival, nearly 16,000 data analysis to tell you where the public opinion
25. Big data report on divorce proceedings in Hangzhou 2015 | kezhi family lawyer team
26. Search 10 | from experience to data - "legal scientist" formation program
27. Ministry of Housing and Urban-rural Development: Five technologies including comprehensively improving informatization level of construction industry and enhancing big data

5.2. Empirical research based on LDA
The content analysis method of 5.1 mainly classifies news headlines. Although it can roughly distinguish the proportion of news quantity, it cannot examine the correlation between the three categories, and the more detailed relationship between news headlines and news content cannot be obtained. Therefore, in Section 5.2, this paper conducts in-depth research on the lexical level of news through LDA algorithm. We summarized all the word participles of 1025 articles to obtain a vocabulary list with a total of 1562,313 words and extracted the 20 topics with the highest frequency, and each topic was described by the 20 words with the highest frequency related to the topic. At the same time, according to the meaning described by the 20 words, they were classified into the three categories of technology, society and business established by the content analysis method in table 2.

| Topic NO. | Topic category | Topic description (20 words) |
|-----------|----------------|------------------------------|
| Topic #0  | Business       | 1. Users, data analysis, product, value, advertising, customer, marketing, behavior, and different, how to, forecast, commodities, recommendations, data mining, purchase, customers, accurate, business, demand, the Internet |
| Topic #1  | Business       | 2. Market, the public, the investment, investors, cars, marketing, sales, 2016, the brand, at present, the capital, view, select, growth, users, increase, in the future, the opportunity to think, to speed up |
| Topic #2  | Technology     | 3. Data, variables, distance, sample, test, distribution, method, Python |
| Topic #3  | Society        | 4. Investment, web site, the design, all, 2015, the public, movies, China city, credit CARDS, respectively, the growth and the country, travel, company, users, industry, and become, market, and content |
| Topic #4  | Technology     | 5. Storage, database, system, Hadoop, processing, data warehouse, computing, query, support, SQL, architecture, platform, implementation, transportation, usage, real-time, structured, user, based, distributed |
| Topic #5  | Technology     | 6. Visualization, use, study, depth, neural networks, charts, tools, design, pictures, different, the machine, a kind of image, computer, language, training, for, vision, network, simple |
| Topic #6  | Technology     | 7. Learning, algorithm, machine, model, method, classification, prediction, neural network, depth, introduction, regression, supervision, base, training, learning, network, data mining, clustering, rules, including |
| Topic #7  | Society        | 8. The financial, government, service, health care, Internet, social, security, open, resources, countries, Banks, innovation, economic, credit, management, enhance, share, use, driving, health |
| Topic #8  | Society        | 9. This is, probably, many, humans, now, no, what, however, has been, in the future, they, not, if, when, oneself, so, in this way, because the artificial intelligence |
Finally, correlation between these 20 topics is generated in Figure 5.

Figure 5 Display of topics relevance extracted by LDA
6. Conclusion
From topics extracted by LDA, we classified 20 highest frequency topics into technology, society and commerce, found that 35% accounted for technology, business accounted for 15%, society accounted for 50%. Compare to the LDA result, the rate of technology, business and society conducted from content analyses is 37%, 22% and 40.7% respectively. The rank of three categories is consistent, but in LDA result, social class accounted for a higher rate, business class account for less rate, technology category accounted for almost the same. The result of LDA analysis reinforce the conclusions of previous content analysis result: the impact of new technologies such as big data and artificial intelligence is more about social thinking than business.

An advantage of LDA over content analysis is that the relevance between topics can be further analyzed. As can be seen from Figure 5.2, topic 1, 5, 6, 10, 14, and 19 are the six most closely related topics in the third quadrant, and they belong to 1 business, 5 technology, 6 technology, 10 technology, 14 society, and 19 technology respectively. In 6 topics, business category up to at least 2 (20 topics in a total of only three topics are affiliated with business), and social class which accounts for highest proportion in content analysis has only 1 topic, other are all categories of technology, it shows the relationship between technology category and business category closer, relatively distant relationship with social categories.

To sum up, at present, people pay more attentions to big data and artificial intelligence technology in social thinking and discussion, and there are fewer commercial cases put into practice, but the relationship between technology and business is closer than that between technology and society.

Acknowledgments
This work was financially supported by the National Natural Science Foundation of China (71832013)

References
[1] Li Shi, Ye Qiang, Li Yijun. Mining the product characteristics and emotional tendency of Chinese online customer comments [J]. Computer application research, 2010, 27 (8)
[2] Johan B, Mao Huina, Zeng Xiaojun. Twitter Mood Predicts the Stock Market [J]. Journal of Computational Science, 2011, 2 (1)
[3] Wei X, Croft W B. Lda – based document Models for ad-hoc retrieval [C]. Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006: 178-185.
[4] Zhang Jianwei, Yan Jianfeng, Liu Xiaosheng, Yang Lu. An LDA algorithm based on dynamic vocabulary [J]. Computer science, 2016, 43 (12)
[5] Hao Jie, Xie Jun, Su Jinqiong, Xu Xinying, Han Xiaoxia. Unsupervised emotion classification based on word weighted LDA algorithm [J]. Journal of intelligent systems, 2016, 11 (4)
[6] Lord Zhang Shu. Comparative Analysis of Online Financial News [J], News Communication, 2014