Measurement of Innovation in the Core Industry of China's Digital Economy: A Text-based Approach

Jie WEI[1], Yuhang SONG[2*] and Yujia REN[1]

[1] Nanjing University, Nanjing, China; winniewei@nju.edu.cn; renyujia95zz@163.com

[2] Northwest University, Xi'an, China; 2350625274@qq.com

* Corresponding author: 2350625274@qq.com

Abstract: With the fourth wave of industrial revolution, the ability to innovate in digital field has become the bottom support of competitive game between countries. This paper focuses on the innovation of the core industry of Chinese digital economy, utilizes the analyst reports of the top 100 enterprises in electronic information industry listed in A-shares, examines the basic characteristics and trends in development of overall innovation and types of China's digital economy by the method of text information mining. Results show that overall innovation of the core industries is currently experiencing rapid progress, many subfields original innovations burst out. Besides, industries incremental innovations and integration of self-developed disruptive technologies has achieved a sense of "curve overtaking". The method of measuring innovation based on text information is a new breakthrough in methodology, which overcomes the limitations of traditional methods, makes up for the technical omissions in measuring unpublished patents or R&D companies.

Keywords: core industry of digital economy; innovation; text information mining; LDA topic model; machine learning

JEL Classification: B41; C55; D21

1. Introduction

The new generation of information technology with artificial intelligence, big data and blockchain as the core is rising strongly, and the digital economy has become an important driving force for global economic development. In recent years, China has seized the opportunity of digital economy development and made new breakthroughs, but the questioning of the current level of innovation in China's digital economy is still a hot issue. Generally speaking, innovation is a multi-dimensional process, and many of its internal processes are difficult to measure. Thus, the research at home and abroad continues on the way to expand the methodology of "innovation measurement" (Haar, 2018). The innovation of electronic information industry as the core industry of digital economy is a typical representative of China's digital economy innovation. At present, China's electronic information industry is taking the initiative to transform the industry to adapt to the development situation at home and abroad, which has shown rapid innovation and high investment growth.

So far, the existing literature mainly focuses on the number of patent applications in electronic information industry (Jia et al., 2021) and research and development input (Liu et

al., 2019; Hao et al., 2020) conducted research, while ignoring the innovation of non-patented and non-R&D enterprises, which may make information omission. In order to further understand the innovation level of electronic information industry, this paper introduces Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) to generate the text topic of analyst report, and uses its good characteristics to model massive heterogeneous text data (Wu et al., 2022). Fully associate the text word aggregation class, effectively measure the topic probability of new documents, and improve the overall accuracy and credibility of the measurement (Omar et al., 2015). Based on this, the difflib function in Python is used to compare the differences between texts, select the best topics and constitute the overall measurement of innovation in electronic information industry with words contained in them. Then further discusses the subdivision type, Original innovation and Incremental innovation, which respectively represent the source of breakthrough and creativity by relying on their own internal forces and the deep mining and development of existing technology (Souto, 2015). Thus excavate the current innovation and development situation of China's digital economy and provide reference for the construction of "digital China" in the future.

Compared with the existing literature, the marginal contributions of this paper are as follows: Firstly, provide a textual description of the innovation level of enterprises. This paper uses Python crawler to conduct data mining on the information contained in industry research reports of the top 100 enterprises to measures the innovation level in the electronic information industry, so as to better explore the innovation connotation under the surface of text words, which can overcome the limitations of traditional innovation measurement methods and further expand the measurement scope. The way to measure innovation makes methodological significance. Secondly, Enlightenment for measuring two different types of innovation. On the basis of measuring China's core industries' overall innovation, this paper further subdivides the forms of innovation, and offer a relatively objective quantitative evaluation index for quantifying.

2. Data and Method

2.1. The Informational Nature of the Analyst Research Report Text

Industry analyst reports are highly informative. Most foreign scholars make use of analyst reports to obtain industry information, evaluate the volatility caused by the analyst reports, and make investment decisions (Wei et al., 2023; Tu, 2022; Roeder et al., 2022). Others use analyst reports to look for factors that influence analyst recommendations (Bouteska & Mili, 2023), measure analyst sentiment (Daudert, 2021) and its possible impact on analyst reports. For example, the forecast deviation of corporate earnings (Jiang et al., 2022), the degree of ethical behavior bias (Brown, 2021), etc. In the context of the reform of information disclosure system, analyst research reports issued by securities institutions have become an important tool for mining and interpreting the company's operation situation and predicting the company's future development direction (Allen et al., 2014). In addition, since the content

of analyst reports includes company innovation, it is possible to obtain the true level of innovation output of enterprises by mining the text of analyst reports.

In recent years, the use of enterprise patents to measure innovation has been widely discussed by scholars. Enterprises usually apply for patents to obtain economic benefits and seek property rights protection, but under some circumstances will not, such as the potential loss of efficiency in applying for patents (Bessen & Maskin, 2009), the cumbersome process of filing patents (Shapiro & Lemley, 2019), which mainly lead by the insufficient protection of creators. In addition, there are still some systematic deviations in patent statistics (Pavitt, 1985), and it is limited to measure the innovation level of enterprises only by measuring the number of patent applications and R&D investment. Therefor the content of analyst research reports can reflect other innovative activities apart from patents and research. For example, Coca-Cola does not apply for patents (Halligan, 2010) to prevent the disclosure of trade secrets, but the value of this secret is well known to analysts. And there are many ways for companies to innovate without filing patents or investing in research and development, like Wal-Mart. The top 500 American enterprise has not applied for patents in the early 1990s, but maintained a high growth rate during this period, and has been developing and improving its supply chain system by using information and communication technology to strengthen its service and brand strengths (Bahramimianrood & Bathaei, 2021).

As the core industry of the digital economy, China's electronic information industry is undergoing multiple transformations of industry, products and technology in recent years. Analysts are focusing on the research and development of new products and technologies in the industry, and the information comments and disclosure of related products will help us analyze the innovation field of this area. In addition, analysts are widely regarded as influential participants in the capital markets (Fogarty & Rogers, 2005) and can influence investors who face significant uncertainties about future market developments through their reports (De Franco et al., 2015; Huang et al., 2014). In the operation of the capital market, analysts tend to predict the future development trend of the industry by describing the company's financial performance, which has a certain authority (Stolowy et al., 2022), and thus are considered as important participants in the capital market (Leins, 2018). Moreover, to a certain extent, analyst reports have the internal information privacy of the industry (Naqvi et al., 2021), and they often have a unique sense of smell in the discovery of new products and technologies in order to gain investors' attention. Therefore, this paper uses the analyst research report to analyze the innovation of the core industry of China's digital economy, deeply digs the connotation of the report text, and obtains the technology field that the core industry of the digital economy is currently booming, as well as the specific types of innovation contained in the industry.

2.2. Sample Selection, Data Source and Data Preprocessing

This paper focuses on the core industry of digital technology in China for the year 2022. Top 100 competitive companies in the electronic information industry which shown at the "World Digital Economy Conference 2022" as research object, and focuses on the companies

with A-share listed stocks as research samples. Based on this, the Wind financial terminal database is used to manually sort out the research reports made by analysts of two securities companies, which are Huatai Securities and Everbright Securities, from 2004 to 2023. Then use Python to scrape the report content to obtain the overall samples, which then are de-duplicated, word segmentation and data cleaning.

In order to reduce the noise that may exist in the above process and ensure the accuracy of the obtained content, this paper pretreats the samples in the following ways: (1) Eliminate the enterprises that have not issued A-shares and are not currently listed among the 100 companies in the above-mentioned industries. (2) Research reports published on WeChat public accounts, research reports without permission, research reports with missing data content and restricted research reports only for whitelist users are excluded, and only all available analyst reports published by Wind Financial Terminal are taken as research samples. (3) Clean the content of the research report, delete the report samples with missing content and the invalid pages that repeatedly appear in the report. (4) Use the stop word list to delete common stop words and other words that may interfere with the results. Finally, 1,114 analyst reports from 76 companies were obtained as research samples.

2.3. Construction of Technical Innovation Vocabulary of Core Industries in China's Digital Economy

In order to obtain the current technological development level of the electronic information industry, and find out the initial measurement standard of industrial business, according to the "Electronic Information Industry Classification Notes (2005-2006)" and "Electronic Information Product Classification Notes" (hereinafter referred to as Notes), to build the technical innovation vocabulary of China's digital technology core industry. China's electronic information industry includes 12 industries, aggregate 46 categories, in which the software industry accounts for 1 industry and 3 categories, and the manufacturing industry accounts for 11 industries and 43 categories. This paper uses the product division standard indicated in the notes to define the overall business scope of the electronic information industry, and constructs the initial basic vocabulary to measure the technological innovation level of the electronic information industry.

Firstly, compare and screen the products to find out the complete range of electronic information industry products; Then, the screened words are simplified and associated to obtain phrases that can be recognized by machines, and deleted redundant items, thus to form a basic vocabulary base for measuring enterprise technological innovation. Finally, based on the vocabulary and the definition of electronic information products, compares the word categories of about 1,500 products in the vocabulary within 12 industries, then explore above two types of innovation (part 3.2.).

3. Measurement of Innovation in Core Industries of China's Digital Economy Based on Text Information

This chapter will use Python software to carry out information mining on the analyst report text, refine and measure the current innovation trend of the electronic information

industry (Figure 1). Firstly, this paper uses the LDA topic model to describe the analyst text, utilizes the pyLDAvis, word cloud and other visual tools to show the measurement results. Secondly, use text similarity measure overall innovation of industry. Then, by constructing a basic vocabulary base to measure technological innovation, using Python to search key texts and other means to collect the industrial technological innovation words contained in the analyst report, the level of original innovation and incremental innovation of the industry is obtained. Based on the data mining results, combined with the actual situation of China's electronic information industry, the current innovation development trend of China's digital economy core industry is obtained, then subdivide the specific innovation types.
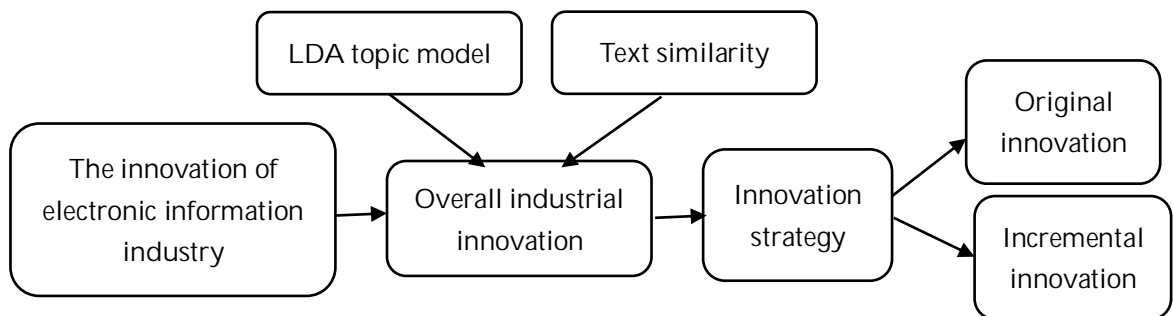

Figure 1. Research frame and process

3.1. Use LDA Topic Model to Measure the Overall Level of Enterprise Innovation

This section mainly uses the LDA topic model to analyze the analyst reports of the top 100 competitive enterprises in China's electronic information industry. Firstly, the optimal number of topics is obtained by constructing the LDA topic model on the crawled reports, and the pyLDAvis visual analysis is carried out. Then, difflib differentiation is compared between the words in the selected topics and the original text, and the topics with the greatest similarity to the innovative textbooks are selected as the optimal topics to measure the overall innovation level, then the display word cloud map is drawn accordingly. Thus, the overall description of the innovation level of the top 100 electronic information industries is obtained.

1. Select the best topic using the LDA topic model

According to the mechanism analysis of the LDA topic model by Blei et al. (2003), it can be seen that LDA model can carry out cross document analysis of given content according to context semantics, effectively extract the document topic, and deduce the probability distribution of the given document topic from the word frequency distribution of the topic, mining and modeling text data (Cao et al., 2009). According to the above definition, in the process of LDA topic analysis, the number of topics is not recovered through data reverse engineering on the basis of fixing itself in advance, but generated by the data itself (Richert & Coelho, 2013), which ensures the reliability of the estimated results. In this paper, Python crawler is first used to crawl text data, and then LDA method is used to construct a dictionary for corpus data and visualize it. By calculating the perplexity curve of LDA model (The et al., 2006), the "elbow method" is used to constantly adjust the value of the
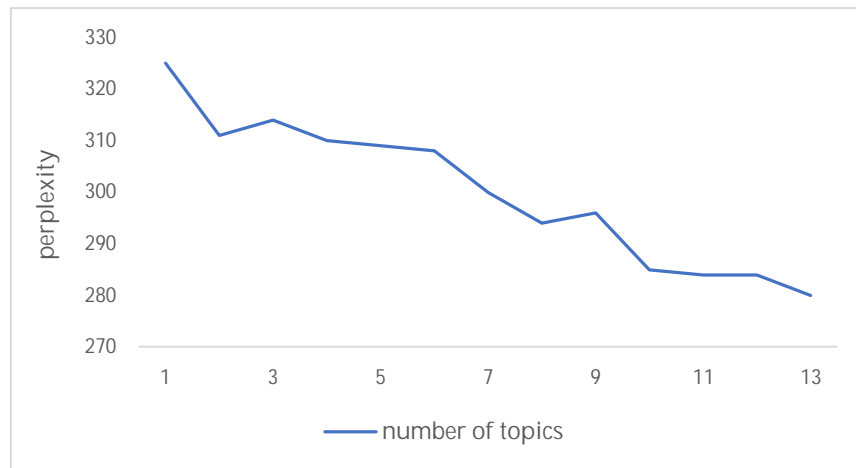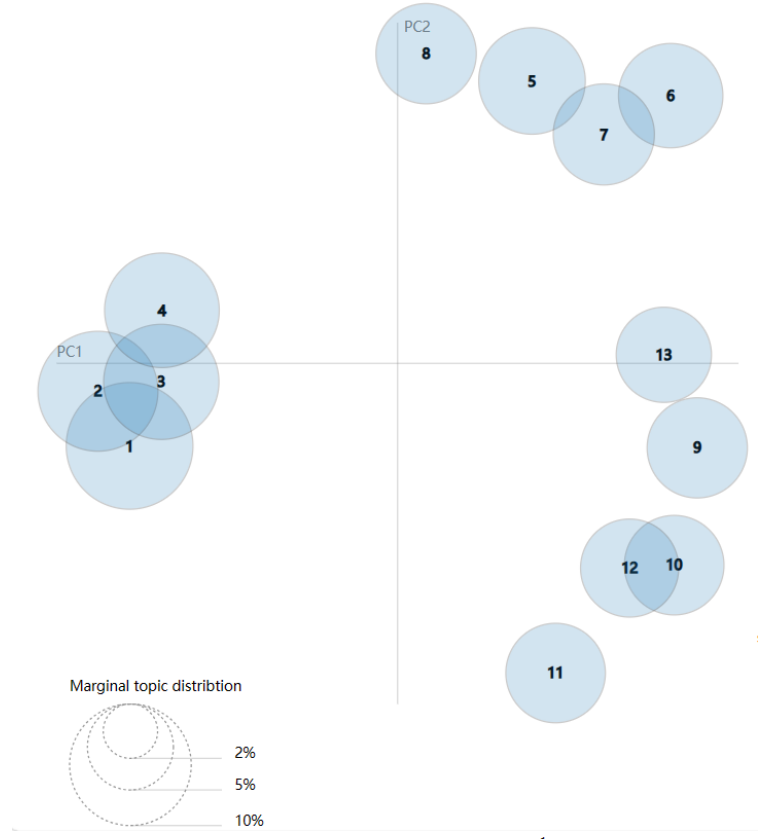
Figure 2. LDA perplexity curve

number of topics k, so that the output topics are kept at a low level of perplexity. Based on this, the optimal number of topics for the target text is determined.

After the above measurement, it is found that the perplexity degree of LDA theme model obtained from the perplexity degree curve is low when the number of topics k=13 (Figure 2). Further, use pyLDAvis to draw theme interaction display graph and verify the suitability of the optimal number of topics obtained from the above perplexity degree curve, then summarize the content of each theme. pyLDAvis is a method that extracts the principal components of text and analyzes them in multiple dimensions. The proximity between subjects is indicated by the distance between subjects. The bubble distance adopts Jensen-Shannon divergence (JSD) distance, and the overlap degree of bubbles indicates the cross-relationship of feature words in the topic (Gottfried et al., 2021). The following Figure 3 shows pyLDAvis images of the above 13 themes. Observing the degree of overlap among the LDA output themes, we can see that the correlation degree among the themes is small and meets the selection criteria of the best theme. Therefore, k=13 is selected as the optimal number of themes output by the LDA theme model.

For each topic obtained by the LDA topic model, the histogram on the right lists the top 30 most relevant words under that topic. Among the 13 topics, we can get the relevant keywords of each topic, and then describe relevant content: In Topic 1, Topic 7 to Topic 13, the LDA model results highlight the future business and operation of the top 100 companies. Here, we output the most relevant 30 words to representative topic 1 to describe the text content (Figure 4, Left). Relevant words express the development of the top 100 companies in the main business: The current revenue and expenses of industries are good, and the gross profit margin has improved overall; Related products are in a rapid development trend, and it is expected that there will be greater breakthrough results in the future; However, in the face of complex domestic and international environment, the industry still has some competition and pressure. The above topics generally indicate that the electronic information industry has significant dividends and returns, in which opportunities and risks coexist. The words in Topic 2 to Topic 6 highlight the general situation of the products of the electronic information industry. The text content related to the representative Topic 2 is output here (Figure 4, Right).

Intertopic Distance Map (via multidimensional scaling)

Marginal topic distribtion

- 2%
- 5%
- 10%

Top-30 Most Salient Terms[1]
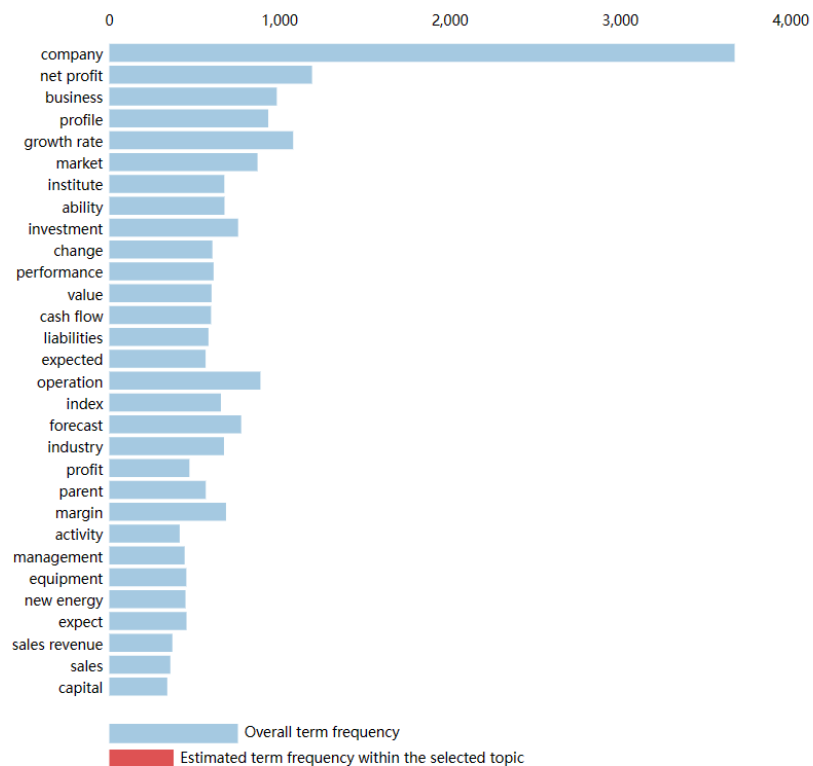
Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. LDA theme topics

Top-30 Most Relevant Terms for Topic 1 (10.9% of tokens)

Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Top-30 Most Relevant Terms for Topic 2 (9.7% of tokens)

Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)
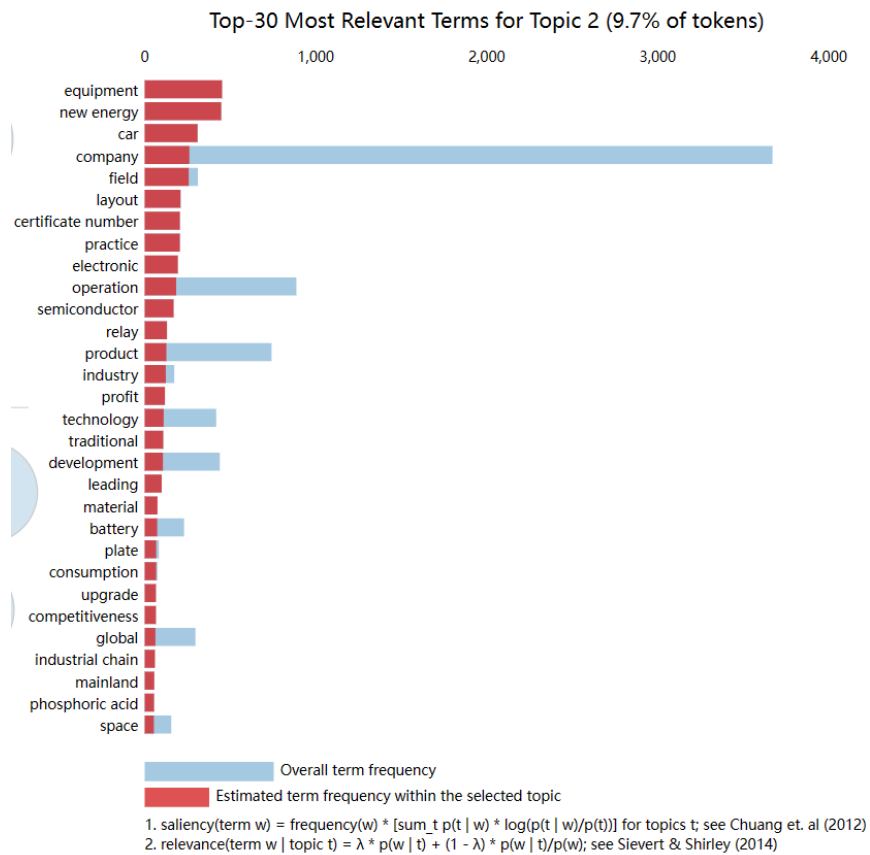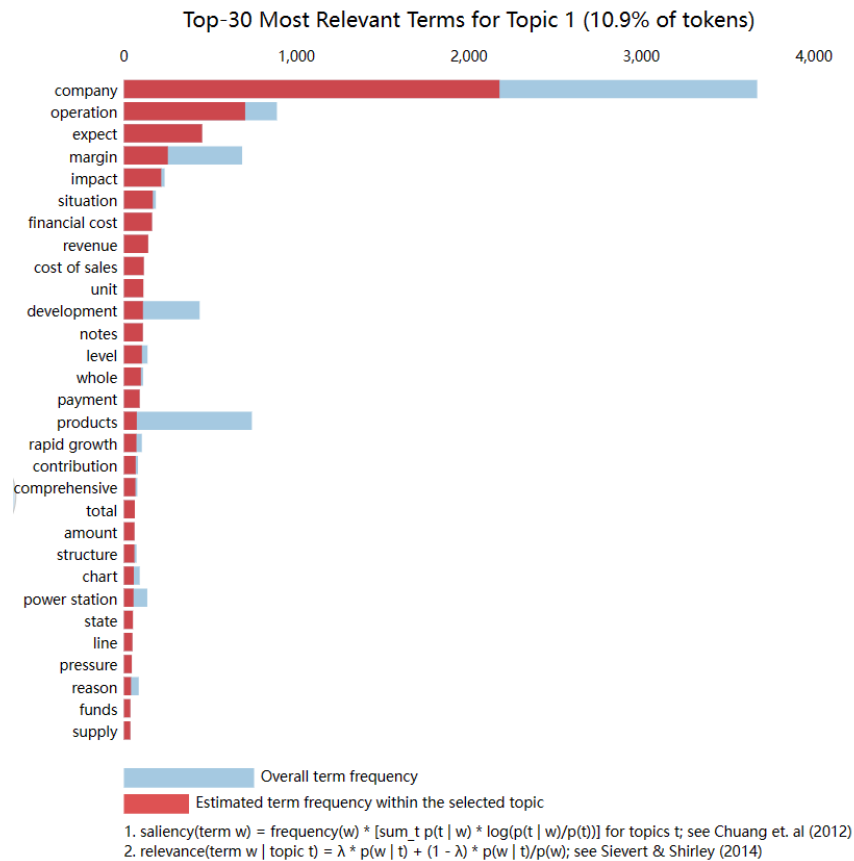
Figure 4. The top-30 LDA representative subject words

These topics describe the related products and equipment in the industry, which mainly mentioned new energy, automotive, semiconductor, relay, flexible industry, smart home, base station, robot, optical fiber, chip, Internet and other hot areas of electronic information industry, which not only reflects the current focus of the development of the electronic information industry, but also indicates the corresponding innovation potential and unlimited innovation possibilities in these subdivided industries.

2. Innovation based on text similarity measurement

In order to measure the overall innovation, we select "Corporate Finance" (Zhu, 2018) as the benchmark innovation textbook to measure the text similarity between subject words. The Python crawler is used to extract the text content in the above standard innovative textbooks, the jieba word segmentation is performed to measure the text word frequency of the textbooks, and the word frequency of the words contained in the book (after weeding out invalid words by using the stop word list) is greater than or equal to 100, and the obtained string represents the original text content and is recorded as A; Then, the top 25 words with the highest word frequency among the 13 topics measured above are taken as representatives of the content of each topic, and the string representing the text content of the topic is obtained and recorded as $B_n, (n = 0,1,...,13)$. The difflib module in Python is used to measure the pair similarity of the above strings, and the topic with the highest similarity to the base text is obtained as a proxy for the overall innovation of this paper.

Combined with the output results of the LDA model, the topic with the highest similarity in difflib is directly related to the factors that describe the overall level of innovation in the company from the text content. Figure 5 shows the word frequency. This topic describes words related to finance of the top 100 enterprises. At present, the industry is constantly changing, and the overall development of the industry is great; The increase in capital, interest rate and sales will bring more profits and investment interest, which will increase the market value of the company. Thus, the current market share of the enterprise has relatively high position in this industry, which means that the company which has large market share will have competitive advantages in the overall industry. In addition, it also indicates that enterprises with high value, high growth and high revenue have more characteristics of innovative enterprises. In China's electronic information industry, the overall strength of the company is still the key factor for enterprises to carry out overall innovation.

Apart from that, as a representative product of the overall innovation of the electronic information industry, smart phones have become the focus of analysts. Domestic mobile phones represented by Huawei, OPPO, vivo, etc. have emerged, and the update iteration of Apple has become a hot topic for analysts to discuss. At the same time, mobile phones as a key product of the electronic information industry, which development has also led to a series of related technology innovation, such as chip research and development, flexible folding screen, as well as China has conquered 5G technology and perfect AI technology, big data and other technology innovation, are reflected in the development process of smart phones.

Figure 5. Theme word cloud

3.2. Measurement of Innovation Types in the Core Industries of China's Digital Economy: Original Innovation or Incremental Innovation

This paper further subdivides and measures the overall innovation types of the core industries of China's digital economy, that is, use the text analysis method to subdivide the original innovation types and incremental innovation types of enterprises in this industry, and then obtain the current development level of these two innovation types of China's electronic information industry.

1. Measurement of original innovation in the core industries of China's digital economy

The original innovation is the first, breakthrough, extremely advanced, and an important starting source of technological innovation and development. Different from the literature that uses the number of patent applications and R&D investment to measure the original innovation of enterprises, this paper defines the new technology or product generated by the original innovation of industry as the technical words that appear in recent years or for the first time in the analyst report, and uses the text analysis method to measure the original innovation of electronic information industry.

This part selects the analyst reports published in the past five years within the above analysis scope, uses Python software to intercept the semantic context of words including "emerging", "development", "exploit" etc., and conducts manual identification and verification of the obtained results to find technical words that meet the definition of original innovation in recent years. Then take this as the measure of the electronic information industry original innovation development level (Table 1).

From the results we can see, the hot and emerging areas of the electronic information industry are integrated circuit industry, new display industry, big data, artificial intelligence, etc., among which household intelligence, chips, 5G, Web of Things and other words have gradually become emerging words in the electronic information industry in recent years.

Firstly, China's fifth generation mobile communication technology (5G) has been in the forefront of the world. As an important part of the 5G industry, words as communication

Table 1. Description of the original innovation in the analyst report (Part)

Target words	Products	Text description
Emerging	Micro and nano, optical communication, semiconductors, power devices	In MEMS and emerging markets, deep silicon etching equipment has been sold in batches to a number of scientific research institutions and production lines, serving many emerging fields such as micro and nano manufacturing, optical communications, compound semiconductors, and power devices……
Development	Intelligent security, AI, Web of Things	Under the trend of security AI, the company proposed the concept of AI Cloud in order to develop the Web of Things industry in the era of artificial intelligence, and defined the concept of AI Cl……
	Optical fiber and cable	Fiber optic cable profits continue to improve, new business development momentum is good……
	Smart phones, AI, 5G, automotive electronics	Downstream innovative applications as smartphones, AI, 5G, and automotive electronics promote the sustainable development……
	Video, intelligent analytics, big data	The five emerging businesses are developing rapidly, and the company is expected to evolve into an intelligent analysis and big data company based on video information……
	Large-size LCD, flexible OLED	Large-size LCD and flexible OLED bring long-term development……
Exploit	Intelligent household appliance	For the development of U+ intelligent cloud platform and UHome OS intelligent home appliance operating system…... The company's vigorous expansion in smart home has been effective……
	Telescopic tube, conical tube	Has developed telescopic tubes, conical tubes, electronic tubes and other innovative products……
	Intelligent hardware, electronic information	Will carry out integrated joint development in intelligent hardware and electronic information core high-end basic components……
Prompt	Integrated circuit chip	Research and development of nanopilot processes…… China's IC chip manufacturing technology is developing rapidly……
New product	defense camera, intelligent cat eye	Also launched such as active defense cameras, intelligent cat eyes, humanoid detection cameras and other new products……
	400G Ethernet, 5G network	The continuous launch of new products such as 400G and 5G... Laying the foundation for future data center evolution to 400G……

equipment, optical fiber and cable, chips, and base stations are increasingly appearing in the text content, and are constantly developing with the innovation of 5G technology. More importantly, breakthroughs have been made in the development of 5G key technologies, and the independent research and development of key core technologies such as 5G chips and mobile operating systems has been successfully achieved. With Huawei's 5G mobile phone, many technologies in China have achieved breakthroughs from 0 to 1.

Secondly, other chip manufacturing fields have also produced fundamental breakthroughs: The analyst report shows that nanodiameter TSV (Through-Silicon-Via) deep silicon etching technology marking the development of China's advanced packaging process, which has promoted the development of linked multiple industries, such as micro and nano manufacturing, optical communications, compound semiconductors, and plays an important role in high-density integration. In addition, Integrated Circuit Chip (IC chips) has achieved 14nm to 7nm level microelectronic manufacturing process, which marks that China's integrated circuits will reduce energy consumption to a greater extent.

Finally, the development of Artificial Intelligence (AI) has promoted multiple fields into the era of intelligence. Firstly, speech recognition AI, is one of the most mature technologies in the field of AI in China, which maintains a leading position in computing power and AI data services. Secondly, AI security with visual recognition, like AI+ home security, AI+ community security and AI+ building security. Finally, AI+ meteorology, like Huawei Cloud Pangea Meteorological large model, proved the superiority of AI in weather forecasting.

In addition, above analyst report still mentioned the original innovation of other key products in the electronic information industry, such as the Web of Things, intelligent robots, new displays, etc., in which the product-related components and core technologies such as printed circuit board (PCB) high-density, filter miniaturization, and lightweight, and dual-camera technology have also been mentioned by analysts repeatedly. It can be seen that the electronic information industry is currently experiencing the overall innovation of the industry and products, the original innovation capacity of the industry has been significantly improved, emerging areas continue to emerge, and the research, development and production of new technologies and new products are constantly in progress.

2. Measurement of incremental innovation in the core industries of China's digital economy

According to the definition of incremental innovation, the incremental innovation in the electronic information industry is mainly reflected in the continuous and incremental innovation of existing products and technologies, from quantitative change to qualitative change, and finally achieve the overall innovation of the industry. Therefore, this paper defines the development of incremental innovation as the high-frequency product terms that appear most frequently in analyst reports of listed companies in the electronic information industry (Bloom et al., 2021). Based on the second section, the vocabulary of the vocabulary database is compared with the original analyst report to obtain the word frequency results of the entire industry's products in the research time frame of this paper, as a measurement of industrial incremental innovation.

As can be seen from the Table 2 results, in the analyst report from 2004 to 2023, products in the electronic information machine, electronic device, communication equipment industry have been mentioned by analysts many times. Conclusions can be drawn that:

Table 2. Incremental Innovation in Technology (Part)

Industrial classification	Product category	Frequency
Electronic information machine	Battery	220
Electronic special materials	Semiconductor	170
Communication equipment	Phone	137
Computer industry	Network	113
Electronic component	Relay	110
Electronic device	LCD	103
Electronic information machine	Optical fiber	90
Electronic device	Integrated circuit	89
Electronic information machine	Wirecable	79
Electronic information machine	Optical cable	78
Communication equipment	Communication device	57
Electronic device	Chip	49

Firstly, "Battery" has the highest frequency in the analyst report text, power batteries represented by lithium-ion batteries and lithium iron phosphate energy storage batteries have continuously achieved innovative breakthroughs, and some new products such as sodium-ion batteries, solid-state battery technology, cobalt-free batteries, and blade batteries have emerged. China's lithium battery technology is achieving the entire industry chain catch-up, breakthrough innovation reached a new height (Gong & Hansen, 2023), becoming a typical representative of the incremental innovation of China's digital core industry.

Secondly, LCD has become a high-frequency term mentioned by analysts in the electronic device industry products, all kinds of innovation in China's electronic display industry are bursting out, and it is currently in the leading position in the world by gradually imitating innovation "curve overtaking". China's display has gone through many stages -- from CRT, LED screen, LED TV screen, LCD, TFT-LCD, PDP plasma, OLED, AMOLED, etc., gradually developed to today's flexible OLED display technology. At present, China's TFT-LCD panel production capacity has ranked first in the world; TFT-LCD panel production line accounts for more than 80% of the world's construction and planning; In addition, China's OLED display technology has gradually become a new breakthrough in the field of screen display.

Finally, the field of Information and communication has achieved domestic substitution, optical fiber and cable development has entered a new stage. Words such as optical fiber, optical cable and wirecable have become high-frequency words in the electronic information industry. At present, China has launched a series of new products and technologies including ultra-low loss optical fiber, ultra-large capacity optical fiber, special optical cable, submarine cable, optical device, optical module, etc., and has achieved innovative breakthroughs in the realization of 5G new optical cables, mass production of 400G and 800G optical modules, and independent research and development of MESH networks.

In addition, the incremental innovation of products such as air conditioners, integrated circuits, and smart hardware chips also plays a non-negligible role in realizing the development and application of technologies such as artificial intelligence, wearable devices, and the Internet of Things (IoT). It is worth noting that among the measurement results of words related to incremental innovation, words related to battery, relay, semiconductor, mobile phone, etc. are consistent with the words that constitute the overall innovation mentioned above. It can be seen that incremental innovation is an important part of the overall innovation. Words related to integrated circuits, fiber optic cables, 5G, and high-definition displays are consistent with original innovations, indicating that the electronic information industry is not only innovating existing technologies in these fields, but also actively developing new products.

3. Sum up

Based on above, the core industry of China's digital economy is currently in a vigorous trend of innovation and development, achieving breakthroughs from 0 to 1 and making technological improvements and innovations in a number of existing fields. However, there are still many risks and challenges: First of all, there is still some gap between China's original

innovation ability and the world's advanced level in some key areas of the digital economy, which is specifically reflected in the "bottleneck" problem of its core technology. Secondly, due to the late start of China's electronic information industry, the innovation of key products and key technologies in some industries still needs to imitate foreign advanced technologies. Finally, there is still a large gap between the R&D investment of China's electronic information industry and that of developed countries, which is due to the scientific and technological innovation in the industry is still in the early stage. Therefore, China still has a long way to go to achieve scientific and technological self-reliance in the field of digital technology.

4. Conclusion

As the world's most active, most dynamic and most permeable field of innovation, the digital economy is becoming an important driving force for economic and social development in today's world. The level of innovation in the core industries of China's digital economy and how to overcome the limitations of traditional methods of measuring innovation are theoretical and practical problems that need to be solved urgently. Therefore, this paper uses the method of text information mining to measure the overall innovation trend and innovation subdivision types of the core industry of China's digital economy. The method of text information mining provides a better measurement method for the innovation measurement of enterprises without patent applications and R&D investment in the study time domain, and it develops a new measurement method in expanding the sample of enterprises including innovation (Bellstam et al., 2020). Using this innovative measurement method to analyze the text content, it is found that: In terms of overall innovation, China's digital core industry is currently showing the characteristics of rapid progress and other original innovations burst out, the incremental innovation and integration of self-developed disruptive technologies has experienced "curve overtaking". However, we cannot ignore the reality that China's digital technology core industry still has limited enterprise technical capacity accumulation, collaborative innovation mechanism and system is still imperfect, and the problem of "stuck neck" in key core areas is still prominent.

Although this paper uses text analysis to measure and analyze the overall innovation and innovation types of the core industries of China's digital economy, it is still insufficient to explore the key factors affecting the innovation of the core industries of the digital economy, and its clearer path of promotion remains to be studied.

Conflict of interest: none.

References

Bahramimianrood, B., & Bathaei, M. (2021). The impact of information technology on knowledge management in the supply chain. *Journal of Social, Management and Tourism Letter, 2021*, 1–11.

Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A text-based analysis of corporate innovation. *Management Science, 67*(7), 4004–4031. https://doi.org/10.1287/mnsc.2020.3682

Bessen, J., & Maskin, E. (2009). Sequential innovation, patents, and imitation. *The RAND Journal of Economics*, *40*(4), 611–635. https://doi.org/10.1111/j.1756-2171.2009.00081.x

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J., & Tahoun, A. (2021). *The Diffusion of Disruptive Technologies* (Report No.1798). Center for Economic Performance. https://cep.lse.ac.uk/pubs/download/dp1798.pdf

Bouteska, A., & Mili, M. (2023). The role of investor sentiment and valuation uncertainty in the changes around analyst recommendations: evidence from US firms. *Journal of Behavioral Finance*, *24*(1), 73–96. https://doi.org/10.1080/15427560.2021.1948853

Brown, T. J. (2021). Ethics, burnout, and reported life and job attitudes among board-certified behavior analysts. *Behavior Analysis: Research and Practice*, *21*(4), 364. https://doi.org/10.1037/bar0000219

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, *230*, 107389. https://doi.org/10.1016/j.knosys.2021.107389

De Franco, G., Hope, O.-K., Vyas, D., & Zhou, Y. (2015). Analyst report readability. *Contemporary Accounting Research*, *32*(1), 76–104. https://doi.org/10.1111/1911-3846.12062

Fogarty, T. J., & Rogers, R. K. (2005). Financial analysts' reports: an extended institutional theory evaluation. *Accounting, Organizations and Society*, *30*(4), 331–356. https://doi.org/10.1016/j.aos.2004.06.003

Gong, H., & Hansen, T. (2023). The rise of China's new energy vehicle lithium-ion battery industry: The coevolution of battery technological innovation systems and policies. *Environmental Innovation and Societal Transitions*, *46*, 100689. https://doi.org/10.1016/j.eist.2022.100689

Gottfried, A., Hartmann, C., & Yates, D. (2021). Mining open government data for business intelligence using data visualization: A two-industry case study. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(4), 1042–1065. https://doi.org/10.3390/jtaer16040059

Haar, P. ter. (2018). Measuring innovation: A state of the science review of existing approaches. *Intangible Capital*, *14*(3), 409–428. https://doi.org/10.3926/ic.1254

Halligan, R. M. (2009). Trade secrets v. patents: the new calculus. *Landslide*, *2*, 10.

Hao, J., Li, C., Yuan, R., Ahmed, M., Khan, M. A., & Oláh, J. (2020). The influence of the knowledge-based network structure hole on enterprise innovation performance: The threshold effect of R&D investment intensity. *Sustainability*, *12*(15), 6155. https://doi.org/10.3390/su12156155

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review*, *89*(6), 2151–2180. https://doi.org/10.2308/accr-50833

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review*, *89*(6), 2151–2180. https://doi.org/10.2308/accr-50833

Jia, W., Xie, Y., Zhao, Y., Yao, K., Shi, H., & Chong, D. (2021). Research on disruptive technology recognition of China's electronic information and communication industry based on patent influence. *Journal of Global Information Management (JGIM)*, *29*(2), 148–165. https://doi.org/10.4018/JGIM.2021030108

Jia-Jia, H., Chunling, L., Runsen, Y., Pervaiz, K., Khan, M. A., & Xiaoran, S. (2022). Dual Innovation Performance through Knowledge-Based Network Structure: Evidence from Electronic Information Industry. *Engineering Economics*, *33*(1), 47–58. https://doi.org/10.5755/j01.ee.33.1.25899

Jiang, X.-F., Xiong, L., Cen, T., Bai, L., Zhao, N., Zhang, J., Zheng, C.-J., & Jiang, T.-Y. (2022). Analyst sentiment and earning forecast bias in financial markets. *Physica A: Statistical Mechanics and Its Applications*, *589*, 126601. https://doi.org/10.1016/j.physa.2021.126601

Leins, S. (2018). *Stories of capitalism: inside the role of financial analysts*. University of Chicago Press.

Liu, D., Chen, T., Liu, X., & Yu, Y. (2019). Do more subsidies promote greater innovation? Evidence from the Chinese electronic manufacturing industry. *Economic Modelling*, *80*, 441–452. https://doi.org/10.1016/j.econmod.2018.11.027

Ministry of Industry and Information Technology of the People's Republic of China. (2023, January 19). *Communications Industry Statistical Bulletin 2022*. https://wap.miit.gov.cn/jgsj/yxj/xxfb/art/2023/art_3f427b68c962460cbe8ebdd754fe7528.html

Naqvi, S. K., Shahzad, F., Rehman, I. U., Qureshi, F., & Laique, U. (2021). Corporate social responsibility performance and information asymmetry: The moderating role of analyst coverage. *Corporate Social Responsibility and Environmental Management*, *28*(6), 1549–1563. https://doi.org/10.1002/csr.2114

Omar, M., On, B.-W., Lee, I., & Choi, G. S. (2015). LDA topics: Representation and evaluation. *Journal of Information Science*, *41*(5), 662–675. https://doi.org/10.1177/0165551515587839

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: Possibilities and problems. *Scientometrics*, *7*(1–2), 77–99. https://doi.org/10.1007/BF02020142

Richert, W., & Coelho, L. P. (2013). *Building Machine Learning Systems with Python* (2nd ed.). Packt Publishing.

Roeder, J., Palmer, M., & Muntermann, J. (2022). Data-driven decision-making in credit risk management: The information value of analyst reports. *Decision Support Systems*, *158*, 113770. https://doi.org/10.1016/j.dss.2022.113770

Shapiro, C., & Lemley, M. A. (2019). The role of antitrust in preventing patent holdup. *University of Pennsylvania Law Review*, *168*(7).

Souto, J. E. (2015). Business model innovation and business concept innovation as the context of incremental innovation and radical innovation. *Tourism Management*, *51*, 142–155. https://doi.org/10.1016/j.tourman.2015.05.017

Stolowy, H., Paugam, L., & Gendron, Y. (2022). Competing for narrative authority in capital markets: Activist short sellers vs. financial analysts. *Accounting, Organizations and Society*, *100*, 101334. https://doi.org/10.1016/j.aos.2022.101334

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

Tu, M. (2022). Analyst Report, Information Transparency and Stock Price Synchronization. *Academic Journal of Business & Management*, *4*(14), 41–45. https://doi.org/10.25236/AJBM.2022.041407

Wei, L., Miao, X., Jing, H., Liu, Z., & Xie, Z. (2023). Bank risk aggregation based on the triple perspectives of bank managers, credit raters, and financial analysts. *Finance Research Letters*, *57*, 104213. https://doi.org/10.1016/j.frl.2023.104213

Wu, F., Xu, W., Lin, C., & Zhang, Y. (2022). Knowledge Trajectories on Public Crisis Management Research from Massive Literature Text Using Topic-Clustered Evolution Extraction. *Mathematics*, *10*(12), 1966. https://doi.org/10.3390/math10121966

Zhu, Y. (2018). *Corporate Finance* (2nd ed.). Fudan University.