Social Media Development and Multi-modal Input for Stock Market Prediction: A Review

Jinshui Huang*
School of Management Science and Engineering
Southwestern University of Finance and Economics
Chengdu, China
huangjinshui341000@163.com

Jun Wang
School of Management Science and Engineering
Southwestern University of Finance and Economics
Chengdu, China
wangjun1987@swufe.edu.cn

Qing Li
Research Institute for Digital Economy and Interdisciplinary Sciences
Southwestern University of Finance and Economics
Chengdu, China
41761780@qq.com

Xiaoman Jin
School of Finance
Hebei University of Economics and Business
Shijiazhuang, China
919885180@qq.com

***Abstract*—In the realm of behavioral finance, the influence of information—particularly that disseminated by media sources—plays a pivotal role in shaping stock market prices. This study categorizes social media into three distinct evolutionary phases, each characterized by its media agency and mode of interaction: authoritative media, grassroots media, and digital interactive media. As we transition into the digital economy era, digital interactive media has emerged as the predominant medium for the dissemination of investment-related information in China. Each media type exhibits unique attributes and varying degrees of impact on the stock market. Given its voluminous information output and dual-subject nature, there is a pressing need for digital interactive media to refine its research approaches and methodologies through a systematic review of existing literature. This paper conducts an exhaustive review of 53 seminal journal articles that focus on stock price forecasting, encompassing all three major phases of media development. It critically evaluates the current body of research, highlighting the focal points and limitations from three perspectives: the evolution of stock market media, the processing of information on media platforms, and the development of multi-source heterogeneous models. Looking forward, the paper advocates for the application of NLP techniques to quantify sentiment and the utilization of multi-source heterogeneous large-scale models.**

***Keywords*—Investor Sentiment Analysis, Digital Interactive Media, Stock Market Analysis, Financial Information Dissemination, Media Evolution in Finance**

I. INTRODUCTION

Information has historically played a pivotal role in influencing stock market dynamics, frequently precipitating notable fluctuations in stock prices. Modern behavioral finance has proven that external information related to enterprises, especially media information, also plays a very important role in the market [10].

Advancements in big data and artificial intelligence have catalyzed a dynamic evolution in the forms of media influencing the stock market. From the form of authoritative news release, to the grassroots information release, and further to the emerging interactive digital platforms. During the Internet's nascent stages, mainstream news media and websites predominantly held sway over the dissemination of information, and investors could only obtain the information through limited channels. Therefore, scholars during this period mainly studied media aware stock movements by analyzing financial information from news [4] [14] [39].

The advent of social media has significantly broadened the scope of internet-based information dissemination. Investors can express their opinions through social media such as Twitter and Weibo. Investors are not only the recipients of market information, but also the producers of market information [6] [22] [24].

Currently, interactive media has emerged as an integral component within the market environment. Among them, the most representative ones are Shanghai Stock Exchange's "e-Hudong" and Shenzhen Stock Exchange's "Hudong Yi". According to statistics from the China Stock Regulatory Commission, the registered users of digital interactive media have reached billion level, gradually becoming a new generation of media factors that affect stock price [41].

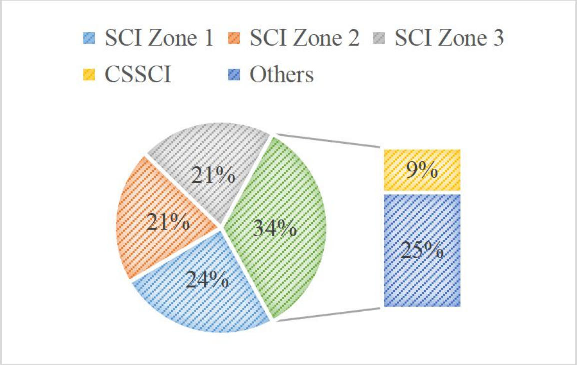

Fig. 1. The Statistics of the SCI Papers Partition

Digital interactive media, established by stock exchanges through the integration of internet technology and new media methodologies, serves as a platform to facilitate standardized, direct, and rapid communication between listed companies and investors. The influence exerted by digital interactive media is multifaceted and complex. On the one hand, it is beneficial for

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

listed companies to disclose information and increase market transparency. On the other hand, it may bring additional noise to the entire market. Regrettably, scholarly discourse on the impact of this new generation media on the stock market remains scant. However, exploring digital interactive media has become particularly urgent and important. This study meticulously reviews 53 leading journal articles on stock price prediction involving three major development stages. It summarizes the focus and shortcomings of existing research from three aspects: stock market media development, media platform information processing, and multi-source heterogeneous model construction thereby paving the way for future explorations in the realm of digital interactive media. The statistics of the SCI partition where 53 papers are located are shown in Fig. 1.

II. Related Work

A. Media Development in Stock Market

Distinct stages in media evolution exhibit varied characteristics of information dissemination, each stage exerting differing mechanisms of influence on the stock market.

- First Stage: Authoritative Media

During the initial stage, control over information release was predominantly in the hands of news media and portal websites. Therefore, during this period, scholars mainly studied the impact through amount of information (news, announcements, financial statements, etc.) or the proportion of emotional words contained in news [4] [13] [39].

Regarding information quantity, Fang and Peress [15] utilized the volume of news from mainstream newspapers as an indicator to analyze the correlation between media news coverage and stock market price fluctuations. Klibanoff, Lamont, and Wizman [19] obtained the number of front page news from 25 national stock markets between 1986 and 1994, and the research results confirmed the significant impact of media information on stock markets. Alanyali, Moat, and Preis [2] collected the number of daily published Financial Times news, and revealed a positive correlation between the number of financial news announcements and the trading volume of company stocks. Meulbroek et al. [29] analyzed the quantity of Dow Jones News and found a direct correlation between news quantity and stock trading volume and returns in the stock market. Chan [7] used the number of news as an independent variable in the econometric model to test the impact of company news on monthly stock returns.

Focusing on the emotional word proportion, Tetlock [39] identified a correlation between the proportion of emotional words in media information and the stock market. The high pessimism of the media indicates that prices would face downward pressure. Li et al. [24] used the proportion of emotional words in media news to calculate public sentiment indicators and fused them with fundamental information to analyze price trends.

However, due to the singularity of the media source, the quantification of media information mainly uses the quantity of information or the proportion of emotional words and maps the relationship between media information and stock volatility through classical econometric linear models.

- Second Stage: Grassroots Media

The advent of Web 2.0 heralded a phase where investors could articulate their views via social media platforms such as stock forums, bars and Weibo. Therefore, the "investors" and "internet users" in the stock market are highly coupled. Discussions, messages and implicit public emotions have a significant impact on investment decisions [5]. Specifically:

In early research, most scholars relied on traditional econometric methods: Das and Chen [9] obtained message board news from Yahoo Finance based on data mining technology and analyzed the correlation between investor sentiment and stock market price. Antweiler and Frank [5] used 1.5 million comment posts on Yahoo Finance and Raging Bull as research samples, proving that the number of discussion posts for specific stocks can significantly affect stock returns. Sabherwal, Sarkar, and Zhang [35] found that the impact of social media information on stock returns has a "reversal feature", with a large number of discussion posts triggering short-term increases in stock prices, but then significantly decreasing. Leung and Ton [21] validated the short-term pull-up effect of social media information on stock prices, and the research results showed that social media information is more significant for small cap stocks in the downward cycle of the market. Meanwhile, Mayew and Venkatachalam [28] used professional voice sentiment analysis software to analyze company conference audio files on social media. The experimental results showed that spokesperson voices contained useful valuable information clues.

Subsequent research witnessed a paradigm shift with scholars employing machine learning methods: Dickinson and Hu [10] found that public sentiment in Twitter has a strong correlation with stock prices by using deep neural networks (DNN). Huang et al. [17] used Convolutional Neural Network (CNN) algorithm to predict the impact of public sentiment on stock prices in Twitter and found that CNN performs better in most cases.

In summary, the escalating complexity of social media texts coupled with advancements in natural language processing technology, scholars have shifted their analysis of emotional dimensions from simple positive and negative judgments to high-dimensional measurements. Scholars have attempted to study advanced machine learning technologies to deeply capture the impact of media information on stock market pricing [52].

- Third Stage: Interactive Media

In the current landscape, the emergence of interactive Q&A communication between investors and listed companies signifies the stock market's transition into its third stage: Interactive Media.

Building upon this, several scholars have embarked on exploring digital interactive media, endeavoring to unravel its impact [53]. For instance, Cen et al. [46] discovered that on 'Hudong Yi', heightened investor attention correlates with reduced stock volatility and liquidity risk. Afterwards, Cen et al. [47] found investor attention can reduce information asymmetry and reduce the cost of corporate equity financing. Ding et al. [48] further found on the "e-Hudong" the improvement of investors' information ability measured by the number of interactive words can significantly reduce the risk of stock price collapse. In addition, Tan et al. [50] used the launch of "Hudong Yi" as an exogenous impact to examine the impact of online interactive platforms on the information efficiency of the stock market from a macro perspective.

In the realm of digital interactive media, the dialogue between investors and listed companies, specifically the Q&A information, is characterized by attributes like temporal delay and diverse, often controversial viewpoints. These are all very important for us to understand the deep mechanism of the impact of digital interactive media on the stock market.

Consequently, research delving into the effects of digital interactive media on the stock market necessitates a departure from previous models. This calls for the development of specialized quantitative methods tailored to Q&A interactive data, alongside the formulation of models based on dual-agency perspectives.

B. *Multiple Processing of Media Source Information*

Unlike scalar data, such as financial transaction figures, media information presents as unstructured textual data, posing unique challenges for analysis. Therefore, how to quantify media information and extract value features from media information has always been a classic and important topic in this field.

Initially, constrained by technological limitations, scholars predominantly employed statistical counting methods for media information analysis. Specifically, Alanyali, Moat, and Preis [2] explored the relationship between daily mentions of listed companies in the Financial Times and the daily trading volume of the company's stock. Preis, Moat, and Stanley [34] provided empirical support for the number of search keywords as a "warning signal" for stock market trends. However, the text content of media information is very complex and the semantics are also very rich. Quantitative methods based on information counting can cause information loss when high-dimensional data is converted into low-dimensional data. Obviously, these methods would lose a large amount of valuable information.

As Natural Language Processing (NLP) has advanced, so too have the techniques for quantifying media information. Scholars have begun to attempt to deeply extract content text by quantifying information into "word vectors", expressing it in the form of subsets of nouns, adjectives, adverbs, and other words. For example, Schumaker and Chen [36] used a weighted method for information retrieval and data mining (TF-IDF) to represent media information to predict the trend of stock price fluctuations after 20 minutes of media information release. Wang et al. [41] characterized media information as a feature vector to assist in financial time series prediction. Akita et al. [1] transformed media information into paragraph vectors and used them together with numerical scalars for stock market price analysis. Wutherich et al. [42] transformed media information published on news websites into word vectors and analyzed the relationship between internet news and daily closing price changes of major stock indices in Asia, Europe, and the United States.

However, the theme of media information content is mainly represented by core vocabulary. There are many irrelevant words in the media information content. Referencing all vocabulary for representation not only reduces the accuracy of the theme expression of media information content, but also introduces a large amount of "noise" information. Subsequently, researchers began exploring the emotional dimensions of media information. Therefore, scholars have begun to pay more attention to the expression of emotions in media information.

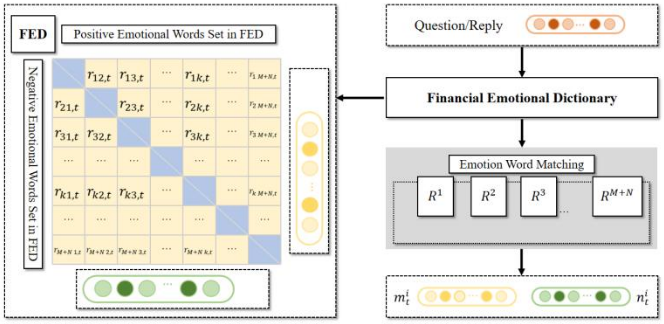
FED
Positive Emotional Words Set in FED
Negative Emotional Words Set in FED
$r_{12,t}$ $r_{13,t}$ … $r_{1k,t}$ … $r_{1\,M+N,t}$
$r_{21,t}$ $r_{23,t}$ … $r_{2k,t}$ … $r_{2\,M+N,t}$
$r_{31,t}$ $r_{32,t}$ … $r_{3k,t}$ … $r_{3\,M+N,t}$
$r_{k1,t}$ $r_{k2,t}$ $r_{k3,t}$ … … $r_{k\,M+N,t}$
$r_{M+N\,1,t}$ $r_{M+N\,2,t}$ $r_{M+N\,3,t}$ … $r_{M+N\,k,t}$ …
Question/Reply
Financial Emotional Dictionary
Emotion Word Matching
R^1
R^2
R^3
R^{M+N}
m_t^i
n_t^i

Fig. 2. Emotional Word Matching Technology (Accepted by PACIS 2023)

Initially, researchers employed rudimentary sentiment analysis methods, based on the proportion of positive and negative sentiment words to represent news texts. For example, Pinto and Asnani [33] innovated an algorithm framework for automatically extracting key phrases from media information, using BP algorithm to train neural networks for predicting the closing price of a given trading day. Tetlock [39] tested the effectiveness of the number of negative words in predicting company stock returns and fundamental trends.

Further, sophisticated NLP techniques have been utilized to assign weights to key vocabulary and sentences, comprehensively distinguish the emotional polarity of news texts. For example, Dickinson and Hu [10] combined N-gram and "word2vec" text representation techniques with random forest classification algorithms to capture the emotions of Tweets, and demonstrate the correlation between these emotions and company stock prices.

Yang, Mo, and Liu [43] established a financial community in the Twitter field and used a new weighted algorithm to construct an important market sentiment index. The experimental results of cross validation showed that the weighted sentiment was more robust in predicting the financial market. Li et al. [24] used noun collections in news to capture fundamental information about a company and reflected investors' biased opinions through emotional vocabulary. In addition, scholars distinguish the emotions of vocabulary by analyzing emotional semantics. Bollen, Mao, and Zeng [6] used two emotion tracking tools (OpinionFinder, GPOMS) to identify the emotional polarity of sentences. The research results indicate that adding specific dimensions of public sentiment can significantly improve the accuracy of DJIA's rise and fall predictions, with an average percentage error reduction of over 6%.

In conclusion, emotional analysis has emerged as a predominant method among scholars for quantifying media information in academic research. Nevertheless, given the diversity and nuances of language, the interpretation of universal emotional lexicons can vary significantly in specialized fields. For example, "bear" and other negative emotional words have become used in the financial field. Therefore, a professional sentiment lexicon for the stock market of digital interactive media is needed in order to improve the accuracy of stock market price prediction.

C. Multi-source Heterogeneous Model

Contemporary research predominantly employs both fundamental financial data (such as stock prices, trading volumes, turnover rates, returns, etc.) and quantified media information (such as quantity, keywords, emotions, etc.) to construct association mapping models. Among them, the main association mapping models used in the study include econometric models, traditional machine learning models, and deep learning models.

- Econometric Model

Econometric models have been extensively utilized to elucidate the causal relationships between diverse influencing factors and stock market dynamics. Among them, the most classic model is the three factor model proposed by Fama French to study the factors influencing the return differences of different stocks [14]. Many scholars also use econometric models to practice market information at a certain time and before as independent variables, with returns or prices as dependent variables. For example, Huang et al. [16] investigated the tone in company press releases and investors' reactions to tone management. They used a logistic regression model to estimate ABTONE, which is positively correlated with the immediate response of stock prices to earnings announcements. Karabulut et al. [19] used the Vector Autoregressive (VAR) framework to study the bidirectional causal relationship between Facebook's National Happiness Index (GNH) and daily stock market activities. In the research on digital interactive media, scholars focus on the number of questions asked [46] [47] [50], timeliness of responses (Zhang Jixun and Han Dongmei, 2015), clarity of responses [53], whether to open online interactive platforms [50], interactive word count [48] [49] are used as independent variables to study the causal relationship between investor attention and stock price. However, a significant limitation of econometric models is their restriction to scalar data inputs, often resulting in valuable information loss during the conversion of high-dimensional data.

- Machine Learning Model

Advancements in computer science have catalyzed the integration of machine learning models into economic research, converting input data from scalars to vectors. For example, models such as decision trees, random forests, support vector machines, and traditional neural networks can use vector data as input to analyze the nonlinear relationships between various influencing factors and the stock market, demonstrating better and more robust analytical performance.

Specifically, Ammann, Frey, and Verhofen [3] used cluster analysis to study the predictive power of articles in the famous German financial newspaper Handelsblatt on the stock market. Liu et al. [25] used the K-means clustering algorithm to classify 293 American companies with official Weibo accounts, revealing the impact of social media indicators on the study of stock return volatility. Luss and D'Aspremont [27] used support vector machines and text as predictive features to predict the intraday price changes of financial assets and improve classification performance. Li et al. [23] used an extended support vector machine (SVR) model to apply regression techniques to predict future stock prices.

Nonetheless, traditional machine learning models frequently exhibit limitations in aligning with complex financial scenarios when dealing with the nonlinear correlation of media effects in the stock market. For example, when the sample data is unevenly distributed, the KNN model only considers the order of k adjacent samples and ignores the actual distance size of the samples [8]. However, at a certain time window, the distribution of quantitative indicators (such as emotional values) of media information samples in the stock market is often imbalanced, leading to a decrease in the ability of KNN models . The Naive Bayesian model assumes that sample attributes are independent of each other, but in the stock market, there is often correlation between media information attributes [18]. In addition, the SVM model is a commonly used model by scholars to analyze the impact of media information on stock market trends [18]. However, the single core SVM model overly relies on data features, resulting in a decrease in the model's performance when dealing with high-dimensional data features. Therefore, scholars have attempted to introduce the Multiple Kernel Learning (MKL) method, which integrates data features and automatically learns weights, taking into account both heterogeneous feature training and kernel function selection issues.

- Deep Learning Model

In fact, there is not only a linear relationship between internet media information and stock market volatility, but also a complex nonlinear correlation. Deep learning models, renowned for their ability to discern intricate, high-dimensional, and implicit patterns of association, offer significant promise.

Illustratively, Ding et al. [12] devised a deep learning approach tailored for event-driven stock market forecasting, which extracts events from media information texts and trains them with a new neural tensor network. The experimental results show that the model can achieve a significant improvement of nearly 6% in the S&P 500 index prediction. Huang et al. [17] used deep neural networks to evaluate and select predictions of Twitter posts' emotions. The results indicate that under the deep network model, Twitter emotions can improve prediction performance. Peng and Jiang [32] applied popular word embedding methods and deep neural networks to predict changes in stock prices using financial news, significantly improving the accuracy of stock predictions in standard financial databases without using historical price information.

In summary, the evolution from econometric to advanced deep learning models reflects a growing sophistication in analyzing the nuanced interplay between media information and stock market behavior.

III. Future Directions

Digital interactive media represents a burgeoning paradigm in the dissemination of stock market information. Its information quantity and dissemination speed are rapidly increasing. At the same time, due to the Q&A interaction mode of digital interactive media, the daily growth of Q&A data has been close to mainstream financial consulting media . Consequently, analyzing the impact mechanisms of digital interactive media on the stock market from a big data perspective becomes imperative. Utilizing advanced NLP technology to automatically obtain and interpret vast quantities of textual information has become one of the important breakthroughs in this field of research.

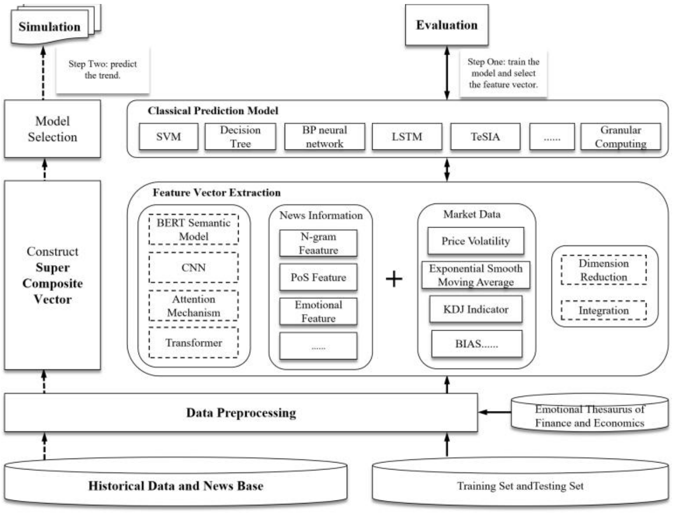

Fig. 3. Framework of Multi-source Heterogeneous Price Prediction Model

In authoritative and grassroots media, the information publisher is often a single subject. The information content in digital interactive media encompasses dual subjects: investors and listed companies. Investors raise questions, and listed companies provide corresponding answers to these questions. Due to the dual-agencies nature, on the one hand, the viewpoints and positions contained in digital interactive media information content are often controversial; On the other hand, there is a certain time interval between the release of each content of the two entities. Hence, research into the impact of digital interactive media on the stock market must transcend the single-subject perspective typical of traditional media. It is necessary to creatively study the impact of the dual subject characteristics of digital interactive media on the stock market.

In traditional research on media effects in the stock market, the independent variable (input variable) is news or social media, and the dependent variable (output variable) is an indicator of stock market (such as cumulative abnormal returns, stock prices, etc.). The correlation mapping between the two mainly relies on classical econometric models or traditional machine learning models. In the context of dual entities in interactive media, this mapping is transformed into a diversified correlation mapping, that is, the content of the dual entities and the impact of their similarities and differences on market returns. Thus, the exploration of a dual agent-based multiple mapping model represents a pivotal breakthrough in understanding the effects of digital interactive media on the stock market.

Acknowledgment

The authors gratefully acknowledge the support of this work by the Fundamental Research Funds for the Central Universities, grant number JBK 2107015. Additionally, the contributions of the Financial Innovation Center of the Southwestern University of Finance and Economics have been invaluable in the realization of this research.

References

[1] Akita, R., Yoshihara, A., Matsubara, T., Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information [C]. In: Proceedings of 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1-6.

[2] Alanyali, M., Moat, H. S., Preis, T. (2013). Quantifying the relationship between financial news and the stock market [J]. Scientific Reports, 3(1), 1-6.

[3] Ammann, M., Frey, R., Verhofen, M. (2014). Do newspaper articles predict aggregate stock returns? [J]. Journal of Behavioral Finance, 15(3), 195-213.

[4] Andersen, T. G., Bollerslev, T., Diebold, F. X., Vega, C. (2007). Real-time price discovery in global stock, bond and foreign exchange markets [J]. Journal of International Economics, 73(2), 251-277.

[5] Antweiler, W., Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards [J]. Journal of Finance, 59(3), 1259-1294.

[6] Bollen, J., Mao, H., Zeng, X. (2011). Twitter mood predicts the stock market [J]. Journal of Computational Science, 2(1), 1-8.

[7] Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines [J]. Journal of Financial Economics, 70(2), 223-260.

[8] Cover, T. M., Hart, P. E. (1967). Nearest neighbour pattern classification [J]. IEEE Transactions in Information Theory, 21-27.

[9] Das, S. R., Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web [J]. Management Science, 53(9), 1375-1388.

[10] Dickinson, B., Hu, W. (2015). Sentiment analysis of investor opinions on twitter [J]. Social Networking, 4(3), 62-71.

[11] Ding, X., Zhang, Y., Liu, T., Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation [C]. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1415-1425.

[12] Ding, X., Zhang, Y., Liu, T., Duan, J. (2015). Deep learning for event-driven stock prediction [C]. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, pp. 2327-2333.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[13] Fama, E. F. (1965). The behavior of stock-market prices [J]. Journal of Business, 38(1), 34-105.

In light of spatial constraints within this publication, it has not been feasible to include the entirety of our references. To ensure comprehensive access and facilitate further scholarly inquiry, we have compiled a complete list of references, which is available for consultation. This exhaustive reference list has been made publicly accessible at the following online repository. We encourage readers to refer to this extended compilation for a more in-depth exploration and research context.

https://pan.baidu.com/s/1gDzXe0MXByedqElxn16_wg?pwd=2024