Research and Analysis on the Application of NLP in Quantitative Investment in the Stock Market

Yinda Liu[1*]

[1]Leicester International Institude, Dalian University of Technology, 116000, Dalian, China

Abstract. The continuous iteration and innovation of Natural Language Processing (NLP) technology offer a technical foundation for quantitative investment in the stock market to break through the inherent limitations of traditional structured data analysis, successfully achieving deep mining, semantic analysis, and quantitative transformation of various unstructured financial texts, such as research reports, financial news, market stock reviews, and company announcements. This paper systematically reviews the domestic and international research findings and empirical conclusions of NLP in the field of quantitative stock investment over the past 3-5 years. Beginning with the core NLP application technologies for customized financial text processing, it focuses on analyzing the technically adaptive strategies and application effects in two key practical scenarios, respectively periodic stock price prediction and stock selection strategy optimization. It further deeply analyses industry controversies and practical limitations in the application of NLP technologies in this area, such as information validity, model interpretability, and factor stability. Combined with the development characteristics of financial markets and technological iteration trends, this paper proposes targeted future optimization directions and research paths, aiming to offer both theoretical sight and practical value for future research, technical implementation, and practical applications at the intersection of NLP and quantitative stock investment.

1 Introduction

Quantitative investment centers on mathematical models and algorithms to complete the systematic and automated execution of investment decisions. Owing to its characteristics of objectivity, discipline, and replicability, it has been regarded as a major developmental trend in the field of stock market quantitative investment. The construction and iteration of traditional models mostly rely on structured data such as price and volume trends, financial indicators, and market trading volume. Such data feature high standardization and suitability for quantitative analysis, and have provided crucial support for model-based decision-making in past investment practices. However, they also suffer from notable limitations in information dimensions: they struggle to effectively capture the valuable information embedded in unstructured text, including market sentiment, policy orientation, industry public opinion, interpretations of corporate announcements, and analyst research report perspectives [1]. Such information is often critical to both short-term fluctuations and medium-to-long-term trends of stock prices, creating an information blind spot in traditional quantitative investment decision-making, so that restricts the prediction accuracy and strategy effectiveness of models.

The rapid iteration and in-depth development of Natural Language Processing (NLP) technologies have provided a key technical pathway to overcome this limitation. Through a series of customized processing procedures, NLP can transform fragmented and non-standardized financial unstructured texts—such as research reports, financial news, stock reviews, corporate announcements, and social media public opinion-into quantifiable feature factors that can be integrated into models, enabling deep mining and value extraction of unstructured information [2]. It has thus become an important breakthrough in remedying the shortcomings of traditional models and driving the optimization and upgrading of quantitative investment models [3].

In the past three to five years, with the widespread application of deep learning in NLP and the growing market demand for diversified information mining in financial markets, scholars and industry practitioners worldwide have conducted extensive theoretical research and empirical analysis on the cross-application of NLP in stock quantitative investment. A comprehensive research framework has gradually taken shape, encompassing customized processing of financial texts and deep semantic feature extraction, scenario adaptation of NLP models and time series prediction models, and integrated modeling that combines NLP features with traditional quantitative factors. In the current technological applications in this field, there are prominent characteristics of refinement, adaptation, and integration. Refinement refers to customized processing workflows developed for the

* Corresponding author: yl845@student.le.ac.uk

specialized nature of financial texts; adaptation refers to designing differentiated model solutions for different scenarios, for example stock price prediction and stock selection strategies; integration represents the deep combination of NLP factors with traditional structured factors. The related research results have performed well in empirical tests in multiple stock markets, including A-shares and U.S. stocks, fully verifying the innovation, feasibility, and practical value of research in this field, and establishing NLP as a research hotspot at the intersection of financial technology and quantitative investment [4].

In the context of this research background and industry landscape, this paper focuses on the application of NLP in quantitative stock investment which is its core research topic. In order to systematically review the relevant technical systems, implementation scenarios, and real effects. Besides that, recognizing the controversies and limitations in current applications, and coming up with accessible methods to solve them. In particularly, this paper first introduce the core technologies of NLP in financial scenarios, then analyze two key application cases, namely stock price prediction and stock selection strategy optimization. Furthermore, it summarizes the problems and challenges existed in practical applications. Finally, it looks ahead to future research directions and find development trends, aiming to provide a reference for both academic research and industrial practice in this field.

2 Core applications of NLP in stock quantitative investment

Financial texts are characterized by strong professionalism, high timeliness, mixed information, and a high proportion of noise, which are significantly different from general daily texts [3]. Therefore, the application of NLP in stock quantitative investment cannot directly adopt general processing flow. By contrast, it requires customized adaptation for financial scenarios, forming a gradually layered technical system specifically for the financial field-from text preprocessing and traditional semantic extraction to deep semantic modeling [4].

2.1 Text preprocessing

As a fundamental step in NLP applications in the financial field, text preprocessing directly determines the effectiveness of subsequent feature extraction. Due to the characteristics of financial texts, Zhang proposed that the core principles of preprocessing are based on the timeliness of information, should focus on the accuracy of professional terminology, and complete text processing through standardized procedures. The process includes three stages. Firstly, cleaning the text to remove noise such as garbled characters, meaningless symbols, and repeated content. Secondly, in financial scenarios. Segment words and tag parts of speech, then use financial industry dictionary to identify professional terms, thereby constructing a customized financial dictionary. Finally, establish a stopword list for the finance field in order to remove mood particles, and other expressions unrelated to investment decisions, thereby focusing on core information as a basis of subsequent semantic analysis [5].

2.2 Traditional semantic feature extraction

Traditional semantic feature extraction focuses on explicit information in financial texts. Its main objective is to transform them into highly explainable quantitative features, so they can be suitable for scenarios that require strong interpretability, for example analyzing short-term trend in stock market [6]. The key methods combine sentiment quantification and topic extraction. On one hand, through combining financial sentiment lexicon methods with traditional machine learning models, the positive, negative, and neutral sentiments in financial texts can be quantified, and it can generate sentiment feature factors at the same time. On the other hand, topic models are used to summarize the themes of texts, so that quantify industry hotspots and policy directions etc. Therefore, the obtained explicit semantic features can be directly integrated into traditional quantitative models [4].

2.3 Deep semantic modeling

The development of deep learning has made deep NLP models popular in current financial text analysis. Its crucial advantage is that it can capture implicit contextual semantics and time features in financial texts, so that deeply extracting textual information. General pre-trained models such as BERT, can identify contextual semantic differences in financial terminology with few mistakes, and solve the problems of semantic deviation when generic models are applied to financial scenarios [7]. Besides that, sequence modeling methods, including LSTM and GRU, perform well in handling time-series financial texts, such as financial news or research reports, as they can capture the timing patterns of information release and semantic correlations [8]. In addition, Cao et al. put attention mechanism into deep NLP models which are used for predicting stock return. It allows the algorithm to focus weights on key information in financial texts and effectively filter out noise information, finally improving the effectiveness of deep semantic feature extraction [9].

3 Core application scenarios of NLP in quantitative stock market investment

The application of NLP techniques in quantitative stock investment has been implemented in core scenarios, mainly focusing on two major directions: stock price prediction and stock selection strategy optimization. Empirical studies in the past three to five years have developed differentiated technically adaptive schemes for different market cycles and analytical dimensions, achieving a deep integration of NLP technology and quantitative investment practices [4].

3.1 Stock price prediction

Stock price prediction constitutes a core component of quantitative investment. Traditional quantitative models are inadequate to interpret the dynamic impacts of textual information on stock prices and fail to capture the linkage between textual information and volume-price data. By integrating text feature quantification with time-series models, NLP technology establishes a multi-period and multi-dimensional stock price prediction system, which effectively improves prediction accuracy [10].

Based on the unique characteristics of intraday and overnight price fluctuations in the Chinese stock market, Wan et al. constructed a targeted multi-period prediction framework [8]. For ultra-short-term and short-term stock price prediction, temporal textual features extracted by deep NLP are combined with models adept at processing sequential data such as LSTM and GRU to capture short-term shocks on stock prices from real-time public opinions and breaking news. For medium- and long-term stock price prediction, financial text fine-tuned pre-trained models are integrated with multiple linear regression and similar methods to mine medium-to-long-term trend information embedded in industry research reports, policy documents and other texts. In another study, He deeply fused sentiment features extracted by NLP with LSTM and achieved accurate short-term stock price prediction in the A-share market through joint modeling at the feature level [10].

In international research, to address the issue of noisy information in financial texts, Cao introduced a hype factor to adjust the features extracted by NLP. This method effectively reduces information bias caused by irrational market speculation and significantly improves the accuracy of stock return prediction [9].

3.2 Stock selection strategy optimization

Traditional quantitative stock selection relies mainly on structured factors such as financial indicators and volume-price data, which leads to problems including factor homogeneity and incomplete information coverage. The introduction of NLP technology has solved this dilemma by establishing a novel multi-dimensional NLP stock selection factor system, realizing the optimization and upgrading of stock selection strategies [11].

Based on financial text analysis, Wu et al. constructed an NLP stock selection factor system covering three dimensions: sentiment, topic, and information quality [11]. Among them, the sentiment factor reflects market sentiment tendencies toward individual stocks or industries; the topic factor captures the degree of matching between industry hotspots and policy orientations; and the information quality factor quantifies the credibility and completeness of research reports and news releases. On the basis of this factor system, researchers integrated NLP factors with traditional quantitative factors such as price-to-earnings ratio and turnover rate to build a new stock selection strategy and conduct historical backtesting. Empirical results show that the stock selection strategy incorporating NLP factors significantly outperforms traditional strategies in core indicators including cumulative return, Sharpe ratio, and maximum drawdown control, which fully validates the optimization value of NLP factors for stock selection strategies [12].

4 Controversies and limitations of NLP applications

Despite remarkable progress achieved by NLP technology in quantitative stock investment and its status as a research hotspot in finances technology, numerous core bottlenecks remain to be addressed when combining theoretical research with market practice. These issues also represent major controversies within the industry, mainly focusing on four aspects: information validity, model interpretability, customization cost, and factor stability [13].

First of all, the effectiveness of information in financial texts is questionable. Financial texts are characterized by short-term timeliness and a high proportion of noisy data. Excessive extraction of textual information can easily lead to overfitting in quantitative models, resulting in excellent performance in historical backtesting, but a significantly decline in strategy stability in actual trading, making continuous profitability difficult to achieve [14].

Secondly, deep NLP models have the problem of insufficient interpretability. Nowadays, mainstream deep NLP models are regarded as "black box models". As for these models, in the process of feature extraction and quantification of financial texts, they lack clear logic to explain the working principle, which makes it difficult to define the internal relationship between the specific text information and investment decisions. This not only hinders the improvement and iteration of the models, but also fails to control risks when making a quantitative investment [15].

Thirdly, it is difficult and costly to process financial texts. The financial industry always constantly updates its specialized terms, and changes dynamically under the influence of policies and global market. Therefore, NLP models need to continuously update its financial corpora. This requires the R&D team to have strong abilities in finance and NLP, at the same time, it also means high modeling and maintenance costs, so that it could limit the adoption of such technology in small and medium-sized quantitative institutions [14].

Finally, there should be more validation about stability of factors based on NLP. Most existing studies rely on backtesting with historical data. However, the stock market has characteristics of concept drift and style rotation, which often weaken the effectiveness of these factors. Since their long-term stability across different time periods and markets has not been widely proven, their reliability in practical applications still needs to test further [13].

5 Future research and application prospects

Based on the current research achievements, industrial controversies and practical limitations of NLP in the field of stock quantitative investment, the future development of this field will center on four core directions: technological optimization, scenario deepening, factor improvement, and data integration. This will promote the in-depth integration of NLP technology and stock quantitative investment, and enhance the theoretical value and practical trading value of the technology. The specific optimization directions are as follows.

Firstly, develop financial-specific NLP models to reduce modeling and maintenance costs. Abandon the simple fine-tuning of general-purpose NLP models, and design fundamental financial-specific NLP models combined with the scenario characteristics of the stock market. Meanwhile, introduce few-shot learning and incremental learning technologies to realize rapid model updates for new terms and new scenarios, significantly reduce the costs of model modeling and subsequent maintenance, and improve its dynamic adaptability to the financial market.

Secondly, integrate explainable artificial intelligence techniques to solve the dilemma of “black-box models”. By combining explainable artificial intelligence with deep NLP models, the specific influence mechanisms of textual information features on stock price prediction and stock selection decisions can be clarified through feature contribution quantification, decision path visualization and other methods. The contribution of textual information to investment decisions can be quantified, realizing models that are interpretable, traceable and tunable, and balancing the accuracy and interpretability of models.

Thirdly, improve the NLP factor system and establish a dynamic effectiveness monitoring mechanism. On the basis of existing sentiment, topic and information quality factors, explore new dimensional factors such as information dissemination speed, credibility of information issuers, and the linkage between textual information and volume-price data, so as to enrich the NLP factor system. Meanwhile, combine with changes in market conditions to maintain the long-term stability of the NLP factor system.

Fourthly, promote multi-source data fusion modeling and build a multimodal quantitative model. Break down the analytical barriers between structured and unstructured data, and construct a multimodal quantitative investment model by integrating multi-source data including NLP textual features, volume-price structured features, financial indicator features and macroeconomic features. This will enable comprehensive capture of market information and make up for information blind spots in a single data dimension.

Fifthly, realize scenario-specific refined applications and enhance the practical value of technology. According to the characteristics of different investment styles (such as value investment, growth investment, short-term trading), different market sectors (large-cap blue chips, small and medium-sized enterprises, STAR Market) and different market cycles (bull market, bear market, volatile market), build differentiated NLP-adapted models to avoid a one-size-fits-all modeling approach, and improve the practical value of NLP technology in various quantitative investment scenarios.

6 Conclusion

NLP technology has opened a new window for analyzing unstructured financial texts in quantitative stock market investment, becoming a core technical means to break through the information limitations of traditional quantitative models. In the past 3–5 years, research in this field has achieved a significant leap from adapting basic technologies to financial scenarios to their implementation of core investment applications, forming customized processing workflows for financial texts, a multi-period stock price prediction system, and a multi-dimensional NLP stock selection factor system. Many empirical studies in different stock markets have verified their optimization effects on quantitative investment models, providing new theoretical and technical support for the innovative development of quantitative investment.

Meanwhile, the current application of NLP in quantitative stock investment still faces key challenges such as disputes over information effectiveness, insufficient interpretability of deep models, untested long-term stability of factors, and high costs of customized modeling, which restrict the large-scale application of the technology in real trading. In the future, with the development of financial-specific NLP models, the integration of explainable artificial intelligence, the advancement of multi-source data fusion modeling, and the improvement of the NLP factor systems, the intersection of NLP and stock quantitative investment will continue to deepen. It will become an important development direction of financial technology, providing more comprehensive and precise support for quantitative investment decisions, and playing an increasingly crucial role in investment decision-making of stock market.

References

1. Z. M. Wang, J. Zhou, J. W. Liang, et al. Financial Public Opinion Crawling and Analysis Based on NLP and Deep Forest. E-Business, (08): 53–54. (2020)
2. T. Meng. Research on the Impact of Social Media on Stock Price Volatility in China. Southwest University of Finance and Economics. (2020)
3. F. Zhang. Analysis on the Impact of Stock Comment Sentiment on BYD Stock Price. East China Jiaotong University. (2023)
4. A. F. Ling, W. Peng, Q. Q. Wang, et al. Advances in the Application of Natural Language Processing in Financial Research. Systems Engineering—Theory & Practice, 44(01): 387–421. (2024)
5. Z. Y. Zhang. Design and Implementation of Financial Announcement Classification and Extraction System Based on NLP. Beijing Jiaotong University. (2019)

6. X. Y. Huang. The Impact of Official Media and Online Self-Media on China's Stock Market. Southwest University of Finance and Economics. (2020)
7. K. S. Desale, Natural Language Processing. In: Concept Drift in Large Language Models. 1st ed. CRC Press, United Kingdom, pp. 59–69. (2025)
8. X. Wan, et al. Do Stock Prices Follow Random Walk over Day and Night? Evidence from Chinese Stock Market. Applied Economics, 56(60): 9117–9120. (2024)
9. Z. Cao, & Geman, H. A Hype-Adjusted Probability Measure for NLP Stock Return Forecasting. Frontiers in Artificial Intelligence, 81527180. (2025)
10. Y. X. He. Research on the Application of Deep Learning NLP Technology and LSTM Model in Quantitative Investment Strategies. Shanghai University of Finance and Economics. (2021)
11. Y. X. Wu. Research and Application of NLP-Based Stock Quantitative Trading System. Taiyuan Normal University. (2024)
12. Y. X. Wu, H. B. Li, G. Z. Hu. Optimization Research on Stock Selection Strategy Based on NLP. Modern Computer, 30(03): 76–82. (2024)
13. X. X. Liu, C. W. Kong. Application Scenarios, Challenges and Development Suggestions of Natural Language Processing in Finance. FinTech Era, 31(05): 23–27. (2023)
14. Q. Qi. Hanwang Technology Claims Leading Position in NLP Business and Receives Inquiry Letter from Shenzhen Stock Exchange, Ending Eight Consecutive Limit-Ups. China Business News, 02-09(A08). (2023)
15. Y. F. Fu, M. Y. Wang. A Review of Deep Learning Applications in Finance. Software Engineering, 25(03): 1–4. (2022)