Dean&Francis
ISSN 2959-6157

Financial Risk Identification Methods in the Context of Big Data

Kaiwen Xiao[1,*]

[1]Reading Academy, Nanjing University of Information Science & Technology, Nanjing, China
*Corresponding author: rv803520@student.reading.ac.uk

Abstract:

This paper comprehensively reviews major research advances in financial risk identification, a field of growing importance due to the increasing frequency of financial disruptions across global markets.It systematically examines two primary methodological approaches: text-based and structured-data-based modeling. Text-based techniques utilize natural language processing and sentiment analysis to extract early risk signals from unstructured sources like news reports, corporate filings, and social media. Conversely, structured data methods employ statistical models and machine learning algorithms—including deep learning and ensemble methods—to identify risk patterns from quantitative financial indicators such as stock volatility, credit ratings, and accounting ratios. Following a detailed synthesis of these approaches, the paper identifies current limitations including data fragmentation and detection delays. It consequently proposes future research directions centered on integrating emerging technologies like blockchain for data integrity and IoT for real-time monitoring, while advocating for greater cross-disciplinary convergence with fields like network science and behavioral economics to develop more robust, adaptive, and holistic risk identification frameworks.

Keywords: Finance; risk identification; big data.

1.Introduction

In recent years, with the further development of global economic integration and market economy, more and more people hope to achieve asset appreciation through investment and financial management. However, the increasingly complex trading environment and endless financial fraud have also deterred many people. Financial transaction risks pose significant threats to wealth management: from the individual level, it may cause property losses; from the macro level of the market, it may even trigger a systemic financial crisis. Faced with such a severe financial transaction risk situation, it is particularly critical to accurately identify financial risks.

According to the investigation report of the Federal Trade Commission of the United States in 2024, the annual fraud losses have risen sharply to $12.5 billion. Data show that investment fraud involves the

highest amount of money in all types of fraud, reaching $5.7 billion, highlighting the urgency of risk management in the field of investment.

Financial risk refers to the uncertainty of the future return of the portfolio caused by the change of financial variables, which covers market risk, credit risk, liquidity risk and other types. As the first line of defense for prevention and control of financial crimes, the importance of financial risk identification is self-evident. By accurately identifying risky financial transactions, it can not only effectively protect personal property security, avoid investors falling into the "pit" and suffer heavy losses, but also help financial institutions to continuously optimize the risk provention and control model. This ability is of vital significance for maintaining the security and order of the financial market and building a strong national security barrier. Against the background of increasingly changeable risk forms and increasingly complex transmission mechanism, it has become an urgent task to seek effective methods to identify financial risks accurately and efficiently.

In the early research of financial risk identification, academia mainly relied on structured financial data for analysis. For example, Altman pioneered the use of a multiple discriminant analysis model (Z-score) to predict the risk of bankruptcy based on financial ratios [1]. With the development of machine learning, support vector machine, random forest and other algorithms are widely used in personal credit evaluation and enterprise default prediction, significantly improving the model's ability to fit nonlinear relationships [2]. At the same time, the breakthrough of natural language processing technology has promoted the research of financial risk identification based on text data, and the research form of using text data to predict financial risk has gradually matured. Loughran and McDonald pointed out that there is a significant bias in the financial context. Based on this, an emotion dictionary suitable for the financial field is constructed, which provides a key analysis tool for this direction [3]. In recent years, researchers have further used big language models such as Bidirectional Encoder Representations from Transformers (BERT) to deeply analyze unstructured texts such as notes to financial reports of listed companies, management discussion and analysis, and financial news, from which key negative information such as "lack of liquidity" and "litigation risk" have been extracted, thus realizing earlier warning of business risks [4,5]. To sum up, the current research on financial risk identification presents two core prediction directions: one is risk prediction based on text data, and the other is risk prediction based on structured financial data.

Starting from the background and definition of financial risk, this paper systematically reviews the current two mainstream risk identification methods based on text data and structured data, and discusses the future research directions on this basis, to provide insights for improving the accuracy and reliability of financial risk identification.

2.Technological Evolution of Two Identification Paradigms

2.1 Dual-track Parallel: Evolution and Integration of Financial Risk Identification Paradigm

Building upon the methodological landscape outlined in the Introduction, this section delves into the evolution of the two predominant paradigms in financial risk identification. It examines their respective trajectories and the emerging trend toward integration. The core question guiding this review is: what are the defining characteristics, correlations, and development trends between these two universal approaches? From the perspective of the development of financial risk identification technology, the evolution of identification technology is not a simple linear substitution, but a rich and diverse change of "dual-track parallel, cross-integration". The two paths use the hard indicators of "data" and the soft information of "text" to identify risks, and eventually move towards synergy.

2.1.1 Structured Data Paradigm: From Linear Discrimination to System Intelligence

The identification method of structured data is to find the relationship between some numerical characteristics and a certain risk state through mathematical models, to think from the perspective of nonlinearity and relevance, and to understand and solve risk problems with the help of technical forces. The use of structured data for financial risk identification has early roots, exemplified by Altman's Z-Score model. Altman used multiple discriminant analysis methods to divide the symmetric enterprise sample into life and death groups, and then constructed a reduced linear function of five financial ratios. This creative idea pioneered the quantitative measurement of the possibility of enterprise bankruptcy, and this achievement also laid the foundation for a series of related studies in the later period. However, its linear assumption and fixed historical perspective are very restrictive. It is difficult for numbers and formulas to describe the complete real world, people begin to try to use some nonlinear structures to describe objects. For example, Barboza et al. reviewed and analyzed different types of machine learning methods, and found that when using ensemble learning methods (such as random forest), the accuracy of the best result obtained by applying 17 financial ratios and cross-validation methods is 86% > Z-Score (currently recognized as the best

ISSN 2959-6157

result is 75% -80%), which is far better than previous research results [2]. This indicates that the focus of research has gradually shifted from "finding the best financial indicators" to "designing better learning algorithms". In recent years, the focus of research has shifted from a single entity to the whole complex system, such as Chen, who is one of the typical representatives of this paper[7]. They use the quantile vector autoregression model to establish the risk spillover network between financial institutions, and use the topological characteristics of the network (betweenness centrality, in-intensity/out-intensity) as new risk indicators, and predict through the machine learning model, believing that the risk is a dynamic process flowing through the network rather than a static risk attribute of a single institution [7]. In order to solve the "black box" problem of some complex machine learning models and improve the reliability of making decisions on machine learning algorithms, the influence degree of each feature on the system risk is quantified by using the SHAP value interpretation model.

The latest progress lies in multi-modal data fusion, RiskLabs, proposed by Cao et al. It is one of the representative works, which contains audio data and text transcription data of profit conference calls, market timing data and news data, and extracts the corresponding features through the corresponding encoder. Then the weighted fusion method is used for integration [7]. This also represents that the previous structured data paradigm is no longer limited to tabular numbers, but is working in a more dimensional and diversified direction, trying to integrate and absorb all kinds of forms, including text.

2.1.2 Text Data Paradigm: From Word Frequency Statistics to Context Perception

The transformation of the text data paradigm is associated with natural language processing, and the deductive process from the outside to the inside and from the shallow to the deep is the beginning and limitation of the dictionary method. Because Loughran and McDonald realized that the general emotional dictionary was "not acclimatized" in the financial field, they created a more specialized financial emotional dictionary and used the proportion of negative words to measure the degree of risk, which was simple but essentially a "word bag model". The context between words and the word order structure are not considered [3]. In addition to the dictionary method, there is another method that directly maps the text to a certain market risk value, in which Kogan et al. Use the "bag-of-words model" and Term Frequency-Inverse Document Frequency (TF-IDF) method to extract text features from 10-K files, and use the principal component analysis method to map the text features in this way to predict the future volatility of stock prices. It also proves that text information can directly reflect market risk, thus breaking away from the category of "emotion" [5]. Pre-training language models has become a paradigm shift. BERT-based models can understand the correct meaning of words in a scenario. By using a large number of texts for pre-training, and then using a small number of financial texts for fine-tuning, such a model, on the one hand, can do document-level risk classification; on the other hand, it can accurately find specific sentences containing certain risks (such as "litigation risk" and "lack of liquidity"), to achieve macro-micro.

The subsequent research tends to be more refined and diversified. One is language-culture matching. For example, Huang Bo et al. embedded the selected Chinese emotion dictionary in the Chinese context and the emotion layer obtained by fine-tuning the BERT model into the calculation model, and fused the text feature value calculated by emotion with the financial data, and found that the text feature value has more unique and over-explanatory information than traditional financial ratios [8]. The second is the further mining and application of the theme model. Liu Chao et al. used the Latent Dirichlet Allocation (LDA) topic model to automatically analyze the text, and obtained some potential topics such as "market risk" and "credit risk"; according to the topic distribution of the document, they measured the degree of risk disclosure by calculating the distance between the document and the topic, and then realized the risk index association between the text topic and the tail with the quantitative value [9].

When the structured data paradigm develops to a certain stage, it provides ideas for data processing and analysis for the text data paradigm, and the development of the text data paradigm promotes the structured data paradigm to develop in the direction of multimodal fusion, thus improving the logical system of "dual-track parallel, cross-fusion".

2.2 Criticism and Insight: Common Limitations and Future Breakthroughs of Current Paradigms

Although some progress has been made in current research, financial risk identification still faces many fundamental challenges. Facing up to these difficulties is the premise of defining the direction of future research. The following three limitations are the analysis of the limitations of the current financial risk identification paradigm from the three different dimensions of database, model characteristics and model application environment, which are interrelated and have their own emphasis, and jointly affect the effect of financial risk identification.

First of all, this kind of data itself has the problem of "true" and "false", "real" and "virtual" are inevitably mixed, "early" and "late" are difficult to be completely consistent; Even if the financial statements are truthfully reported, it is also a "description" of the current situation of the enterprise, and there will inevitably be factors of delay, whitewash and even fraud. No matter how good the model is, the foundation of the model is not solid in such a "tall building on the sand". Moreover, models are based on past historical information to predict, unable to cope with sudden geopolitical risks, "black swan" events, or the impact of new technologies, which can easily lead to "empty window" misjudgments.

Secondly, it is difficult for complex models to be interpretable. RiskLabs, deep integration models, and so on can continue to improve prediction accuracy, but their decisions are more opaque. In applications such as finance, which require strong attribution and strong responsibility, the "black box" model that cannot be explained, no matter how accurate, cannot be recognized by supervisors or wind control personnel.

Thirdly, there is a "game effect" between the model and the actors. When a risk model based on text sentiment analysis is widely used, managers may deliberately remove the negative information they have learned from the report text and intentionally use more neutral wording to cover up the existing risks. As a result, the signals that the risk model itself relies on will gradually "fail", so the game model will become a dynamic game rather than a static data mining.

3.Future Research Directions for Identifying Financial Risks

Through combing the existing literature, can find that the current financial risk identification methods mainly focus on text analysis and large-scale financial data modeling. However, these methods still have some limitations. Based on this, this paper attempts to explore the possible future research directions as follows.

3.1 Combination of Financial Risk Identification Technology with blockchain and Internet of Things

The data stored in blockchain is difficult to modify. The unique chain structure and encryption technology can ensure the security and authenticity of data and prevent malicious modification. The use of blockchain technology can also solve the core risks in traditional supply chain finance-information asymmetry and transaction authenticity risk [10]. The application of Internet of Things technology enables us to obtain more effective, more real and more accurate data.

3.2 Interdisciplinary Integration

The main observation objects of traditional financial risk identification, whether it is text data related to finance or digital data directly related to finance, can not be separated from the impact of some non-financial data. Traditional financial models assume that "people are rational", but there are always more or less deviations in real life. Interdisciplinary integration can effectively compensate for this. For example, the introduction of some sociological and psychological behavioral analysis frameworks into the model of identifying financial risks can predict the results of human factors to a certain extent [11].

4. Conclusion

This article, through a systematic examination and comparative analysis of the two research methods, reveals their respective internal logic, applicable scenarios and limitations. On this basis, this study further proposes and demonstrates "dual-track integration" as a feasible path for methodological innovation and deepening.

Firstly, text-based analysis methods are adept at capturing subtle meaning flows, context-dependent narrative logics, and complex subjective constructions, and possess irreplaceable value in terms of understanding depth, context restoration, and meaning interpretation. However, its conclusions are often limited by the interpretive boundaries of the materials and the interpretive perspectives of the researchers, and face challenges in terms of universality and comparability.

Secondly, methods based on structured data enhance the systematicness and verification potential of research through quantifiable and comparable variables and models, and are particularly suitable for revealing macro trends, structural correlations, and probability patterns. However, the potential risks of context deembedding and meaning flattening also require us to remain sensitive to the social and cultural context behind the data while pursuing precision.

This study holds that a truly persuasive research strategy is not to choose between the two, but to move towards a conscious "dual-track integration". This means that in the early stage of problem design, the depth of the text and the breadth of the data should be comprehensively considered. During the analysis process, enable qualitative insights and quantitative evidence to communicate with each other and correct one another. When deriving conclusions, both the richness of meaning and the rigor of reasoning should be taken into account. This integration

ISSN 2959-6157

is not merely a simple juxtaposition or supplement, but aims to achieve methodological synergy - enabling macro trends to be annotated by micro experiences and allowing individual cases to reveal their universal significance through structural comparisons.

In conclusion, the "dual-track integration" represents an integrative and reflective research approach, which requires us to go beyond the methodical disputes and shift towards a problem-centered, flexible and diverse method adaptation. Future research can build on this foundation to further explore the specific operational framework, verification conditions, and application potential of integration in interdisciplinary scenarios, thereby promoting the continuous deepening of methodological awareness and practice.

References

[1] Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 1968, 23(4): 589-609.

[2] Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction. Expert Systems with Applications, 2017, 83: 405-417.

[3] Loughran T, & McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 2021, 66(1): 35-65.

[4] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[5] Kogan S, Levin D, Routledge B R, et al. Predicting risk from financial reports with regression. Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics, 2009: 272-280.

[6] Cao Y, Chen Z, Kumar P, et al. RiskLabs: Predicting Financial Risk Using Large Language Model based on Multimodal and Multi-Sources Data. Proceedings of the International Workshop on Multimodal Financial Foundation Models, 2024.

[7] Chen Y, Wang G J, Zhu Y, et al. Identifying systemic risk drivers of FinTech and traditional financial institutions: machine learning-based prediction and interpretation. The European Journal of Finance, 2024, 30(18): 2157-2190.

[8] Liu C, Qian C. Dynamic Identification and Measurement of Tail Risk Influencing Factors Driven by Text Data: An Empirical Study Based on Chinese Financial Institutions. Journal of Industrial Engineering and Engineering Management, 2025, 39(6): 16-34.

[9] Huang B, Yao X, Luo Y Q, et al. Improving financial distress prediction using textual sentiment of annual reports. Annals of Operations Research, 2023, 330: 457-484.

[10] Fan L, Wu X, & Li Z. Blockchain-based supply chain finance: Mitigating risk and improving efficiency. Production and Operations Management, 2022, 31(10): 3797-3814.

[11] Barberis N, & Thaler R. Investor psychology and asset pricing. In: Constantinides G M, Harris M, & Stulz R M, eds. Handbook of the Economics of Finance. Elsevier, 2003, 1: 1053-1128.