Proceedings of ICMRED 2026 symposium: The Future of Work: Strategy, Workforce Transformation, and Organizational Renewal
DOI: 10.54254/2754-1169/2026.LD33165

Financial Risk Identification of Listed Companies under Multimodal Information: Default Prediction Analysis Based on Light GBM and Z-score Methods

Linlin Dai[1*], Zhilin Xiao[2]

[1]*School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou, China*
[2]*School of Cyber Science and Technology, Sun Yat-Sen University, Shenzhen, China*
**Corresponeding Author. Email: aobin20231@outlook.com*

Abstract. In the current era, corporate bond default incidents occur frequently, and the financial sector has been focusing on how to accurately identify potential default risks of enterprises. Most traditional default prediction methods rely on financial indicator data, which have certain explanatory power, but they are still subject to factors such as information lag and susceptibility to manipulation, leading to limitations in prediction effectiveness. This paper introduces the Management's Discussion and Analysis (MD&A) textual information and integrates it with financial data to construct a default prediction model. The study selects annual report data from 2021 to 2024 of 67 A-share listed companies in our country, generating panel data with 251 observations. This article chooses Light GBM as main algorithm, while financial indicators such as the debt-to-asset ratio and net profit rate, are selected to depict the financial condition of enterprises. From textual data, features such as sentiment orientation and forward-looking descriptions are extracted, by using Bert semantics. This features are then formed into a feature system through dimensionality reduction. The results shows that the model possesses strong discriminatory capability and good identification of high-risk enterprises and a low misjudgment ratio. At the same time it is discovered that forward-looking information contributes to enhancing the accuracy and forward-looking nature of default risk identification, which proves the supplementary value of the text information.

Keywords: Multimodal data Default risk, Light GBM, MD&A text, Machine learning

1. Introduction

1.1. Research background

The continuous expansion of the bond market has provided enterprises with more financing channels, but it has also led to frequent occurrences of default events. In previous studies it is indicated that whether the corporate will break the contract is related to macroeconomic factors, industry development, and the enterprises' own operational conditions [1]. Traditional default prediction methods, like Z-score model, rely on financial ratios to measure bankruptcy risk [2].

DOI: 10.54254/2754-1169/2026.LD33165

These data only reflect past events and are vulnerable to interference from accounting treatments, making the effect of prediction quite limited.

1.2. The incremental contribution of MD&A annual report text

As an crucial part of annual reports, the Management's Discussion and Analysis (known as MD&A) offers greater forward-looking insights, for containing a wealth of operational explanations and judgments about future development. The natural language processing (NLP)techniques also make it possible to delve implicit information from the text.

1.3. Research purpose and significance

Based on the above background, this paper aims to combine the textual features in MD&A with traditional financial data to construct a default risk prediction model based on Light GBM. The effects of the textual information on the timeliness and accuracy in the prediction are also verified when the model's performance is evaluated.

1.4. Literature review

Although existing default prediction methods have already contained Models such as logistics regression, discriminant analysis, support vector machines, or machine learning approaches, most of them rely on structured data and overlook the potential value of textual data [3].Studies have shown that textual features in MD&A can enhance the accuracy of default prediction [4]. Generally, small and medium-sized enterprises will be classified to be high default risk if they have strategic positioning deviations or excessively high equity pledge ratios [5]. Default factors involve internal governance, external environment, and macroeconomic condition. Textual information can effectively make up the deficiency of financial data and eventually reveal potential risks of enterprises [6].

Multimodal data fusion methods have gradually gained attention in recent years. Concentrating on the difference between these two types of features, a multimodal prediction model named STMA is proposed. The study found that such information indeed significantly improves the accuracy of default prediction, thereby confirming the supplementary value of textual information [7].

2. Research methods and design

2.1. Data sources and sample selection

The financial data for this study were sourced from the CSMAR database, while the textual data from the MD&A section was extracted from corporate annual reports. The study only retained those enterprises that consistently disclosed their annual reports from 2021 to 2024, the financial firms are excluded to avoid interfering the research findings. Ultimately, 67 listed company on China's A-share market were selected as the research subjects.

In Figure 1, it is shown that the sample covers 12 major industries, with the combined 70% proportion of industries such as pharmaceuticals and biology, real estate, electronics, automobiles, and computers. It reflects the characteristics of industries with relatively comprehensive information disclosure in the current market and highlights the differences in risk levels across various industries. The data were eventually organized into a panel dataset containing 252 observations.

DOI: 10.54254/2754-1169/2026.LD33165

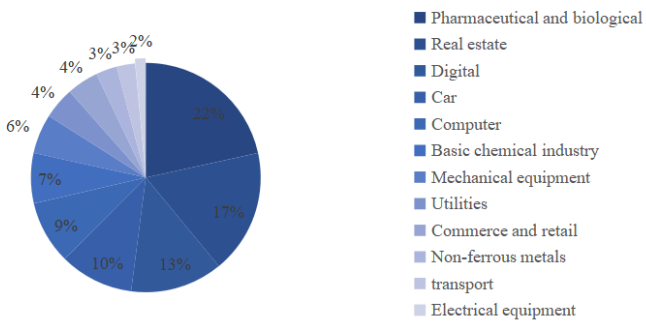

Figure 1. Number of observations

2.2. Definition of default variables

When a company is about to defaults, it will typically undergo a continuous deterioration in its financial condition [8]. Based on this theory, this study uses "financial distress" as a proxy indicator —a company is regarded as being in financial distress when it meets any of the following conditions: The debt-to-asset ratio exceeds 50%;The return on total assets (ROA) is below -10% for two consecutive years; The cash flow from operating activities is negative, and the debt-to-asset ratio exceeds 90%;The Altman Z-score falls below 1.23 (the bankruptcy range).

To capture the rising trend in a company's leverage level earlier and enhance the sensitivity of predictive warnings, the debt-to-asset ratio as a distress indicator is set at 50%. During robustness analysis in this later study, the adjustment of threshold to 70% or 80% lead to minimal sample changes, hardly affecting the conclusions.

The logic of determining financial distress by satisfying any one of the four conditions also matters.it enables the model to more sensitively identify high-risk situations, and comprehensively assesses a company's risk status through multi-dimensional indicators, in terms of solvency, profitability, and liquidity.

2.3. Feature construction

This paper selects variables such as the debt-to-asset ratio and current ratio as financial indicators.

From the MD&A, indicators such as sentiment tendency, degree of uncertainty, and forward-looking expressions content are extracted to be textual features, counting the frequency of vocabulary related to risk, default, and litigation.

The model among them is the ratio of positive word frequency to negative word frequency to construct sentiment tendency, the frequency of such keywords as anticipate, expect, likely and similar is the key to assessing the forward-looking expressions and risk keywords are determined by the appearances of such words as default, litigation, and liquidity risk.

To vectorize the text, the BERT model is employed by capturing the contextual information with its bidirectional Transformer architecture, which is more accurate in semantic measurement. Previous research has shown that the BERT model is more effective than traditional word frequency statistical models in determining whether managers are engaging in tricks (or hiding problems or inflating performance) and predicting the future performance of the company and whether the stock price is going to pop up and the existence of a problem will be known earlier and more accurately [9]. The BERT model is used to obtain MD&A textual features in this paper. The compressions of

DOI: 10.54254/2754-1169/2026.LD33165

high-dimensional features with dimensionality reduction without semantic integrity loss create a comprehensive feature system comprising of 18 variables.

2.4. Model building

Light GBM is more efficient and fits better because it uses histogram-based binning and a leaf-wise growth plan. The grid search and cross-validation are used to obtain model parameters: it capped the number of leaf nodes introduced to 12, and the depth of a tree to 3, to ensure that the nonlinear relationships between variables were captured and that the complexity of models was controlled. With The learning rate 0.02, and the number of rounds 300, the model achieves stable on vergence and prevents overfitting due to a high learning rate.

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

L represents the loss function, and $\Omega(\mathrm{f_k})$ is the regularization term for model complexity, which is used to prevent overfitting.

3. Results

3.1. Cross-validation results

To determine the stability and the generalization performance of the model in various time periods. This paper tests the Light GBM model by holding retraining the model on yearly basis. The model uses the previous year data to forecast the coming year and then it rolls and rolls.

The data is broken down into three distinct sections in order to conduct the analysis as illustrated in Table 1. The AUC remain high across all rounds, indicating the model's strong capability in distinguishing high-risk from other enterprises.

Although the fluctuation in each component is minimal, no major drop in performance is based, which proves that the effectiveness of the model in identifying is comparatively constant in any location. It is notable that the accuracy of the second fold is 100% due to the relatively large percentage of high-risk samples of the test set of the second fold, The model has produced no false positives in this window. It is a manifestation of the excellent discriminatory power of the model in certain time intervals. This drastic outcome also implies that the sample distribution can have an impact on the evaluation indicators. Recall rate in this period is rather high, which implies that this model is sensitive to identifying high-risk enterprises and reducing the risk of missing potentially risky samples.

Table 1. Cross validation results

	AUC	Accuracy	Recall	F1
1	0.9234	79.41%	93.10%	0.8571
2	0.9400	100.00%	76.32%	0.8657
3	0.9647	87.50%	94.59%	0.9091

3.2. Comparison of models with different feature combinations

To compare the predictive power of financial information with textual information for corporate default risk, this paper prepares two prediction strategies: a pure financial model and a fused model

DOI: 10.54254/2754-1169/2026.LD33165

combining financial features with textual. The stability and performance of each model are evaluated through five random partitioning experiments.

3.2.1. Model performance comparison

The Table 2 shows the performance comparison between the two types of models presented. Fused model presents the most significant growth by 97.03% and 118.33% in precision respectively. This indicates that the introduction of textual information can effectively correct the risk assessment biases of financial models and thus reduce the false positive rate. While maintaining a relatively high recall rate, the fused model achieves a substantial increase in the F1 score, suggesting that the model strikes a better balance between controlling false positives and false negatives.

Table 2. Performance comparison of three models

Model category	Accuracy	Precision	Recall	F1	AUC
pure financial model	85.49%	34.68%	28.57%	0.2915	0.8636
Fused model	87.45%	68.33%	45.71%	0.4871	0.8701
Improvement	+2.29%	+97.03%	+60.00%	+67.08%	+0.75%

3.3. Test set confusion matrix

The classification results of the test set are shown in Table 3. The model correctly identified 34 high-risk enterprises and 17 low-risk enterprises totally, demonstrating a relatively favourable overall classification performance.2 low-risk enterprises misclassified as high-risk and 4 high-risk enterprises were undetected, showing that the model exhibits strong detection capability for high-risk enterprises with its high recall rate, which holds practical significance for risk early-warning scenarios. The relatively small number of false positives also indicates that the model also performs robustly in avoiding excessive warnings.

Table 3. Test set confusion matrix

	Predict high - risk situations/events	Predict know - risk situations/events	Sum
Actual high - risk	34	4	38
Actual low - risk	2	17	19
Sum	36	21	57

3.4. Analysis of feature significance

Table 4 presents the results of feature importance ranking. The model mainly relies on financial variables for risk assessment. The debt-to-asset ratio makes the most outstanding contribution, significantly higher than other indicators, indicating the crucial role of company's leverage level in the risk identification process. Indicators such as net profit margin, return on total assets, and current ratio also carry relatively high weights. These variables reflect a company's profitability and liquidity status from different perspectives and are of great significance for risk assessment.

It should be noted that the relatively low importance of textual features is not contradictory to their value, which is not manifested as independent predictive ability but rather through an interaction effect with financial indicators to exert value indirectly. To be more specific for samples

DOI: 10.54254/2754-1169/2026.LD33165

where financial indicators are ambiguous or in a critical state, textual information can provide additional risk signals, correct the mistake in financial judgments, and thereby enhance overall classification performance. Although this "marginal correction" effect is not directly evident in the feature importance ranking, it is validated as the performance of the fused model improve, compared to the pure financial model, being reasonable in this paper.

3.5. Analysis of false positives and false negatives

This paper further analyzes the false positives and false negatives generated by the model. Based on the test results in Table 4, there are 2 false positive samples and 4 false negative samples, with the overall misjudgment rate at a relatively low level.

The false positive samples were not identified as being in financial distress. It is notable that these samples generally already exhibited a certain degree of risk characteristics, such as relatively high debt-to-asset levels or fluctuations in profitability. This may imply that although these companies have not defaulted yet, they are close to doing so. The model simply made an early default judgment on these companies. Such results prove that the model to be forward-looking, to some extent, and to capture potential risk signals in advance. The risk warnings or expressions of uncertainty in the text may also have an impact on the model's judgments.

Further analysis of the MD&A (Management's Discussion and Analysis) text features of the false negative samples reveals that the sentiment tendency scores of these samples are generally higher than those of high-risk companies in the same industry. They have a lower frequency of forward-looking expressions and a significantly lower number of risk keywords. This characteristic is relatively consistent with the previous conclusion that "distressed companies tend to use a more positive tone to cover up problems" [10]. Some companies may also try to conceal potential risks by manipulating the tone of the text even when their financial indicators have not yet significantly deteriorated. The "lies" in companies' annual reports reduce the model's identification sensitivity. These misjudgments reflect the combined influence of the lagging nature of financial information and text expression biases. There is no significant contradiction with the conclusions of this study.

Table 4. Feature importance ranking

sort	feature	importance	category
1	debt-to-asset ratio	8	finance
2	net profit margin	3	finance
3	ROA	2	finance
3	current ratio	2	finance
5	Forward-looking	1	text

4. Conclusion

Using Chinese A-share listed companies as sample, this paper constructs a prediction model with structured data and text data based on Light GBM, and evaluates the model performance through cross-validation of time series. The results show that this model has good predictive ability and stability in identifying potential risks of enterprises, and can accurately distinguish high-risk.

During further analysis, this paper found that structured data still plays a leading role in the model, while MD&A provides extra information. As the structured data may be fuzzy and in the critical state, it can correct this situation and avoid predict mistake, making up for the lag of

DOI: 10.54254/2754-1169/2026.LD33165

structured data and its vulnerability to accounting treatment. The integration of the two types of data enables the model to depict the enterprises more comprehensively, enhancing the ability to identify risks, and verifies the value of multimodal methods in default prediction.

By introducing MD&A and using Light GBM, this paper constructs a multimodal prediction framework, which can verify the value of text information. It also applies BERT process including text representation and dimensionality reduction processing in dealing with the complexity of high-dimensional text features.

The model in this paper can provide detection tools for regulatory authorities, and be served as quantitative references for investors' credit risk assessment.

Despite the exploration of multimodal default prediction, it should be pointed out that there are still limitations in this paper, needing to be further improved.

It is relatively limited when it comes to the research sample selection and time span. Meanwhile the industry distribution has certain concentration, may affect the generalization ability of the model. The research in the future can enhance the robustness of the results by expanding the sample range and extending the time span, adopting industry-stratified sampling or introducing industry-fixed effects. It is undeniable that this article has BERT to extract text semantic information, but the mining for deep semantic feature remains limited. More complex language models can be integrated to further enhance the ability to expressive text features in terms of contextual relations, syntactic structures, and implicit emotions. Light GBM performance, to some extent, depend on the parameters setting, making the classification threshold sensitive. It is expected to adapt model fusion or ensemble learning methods to better reflect the evolution process of enterprise risks. This article has not yet used independent external samples for inspection. hoping that the stability and prospective nature of the model can be deeper evaluated through rolling window prediction or cross-sample testing, improving the reliability of the research conclusions.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Huang, Y. (2024). A Study of Factors Influencing Corporate Debt Default. Highlights in Business, Economics and Management, 40, 23–28.

[2] Bootorabi, F., Haapasalo, J., Smith, E., Haapasalo, H. and Parkkila, S. (2011) Carbonic Anhydrase VII—A Potential Prognostic Marker in Gliomas. Health, 3, 6-12.

[3] Zheng Dachuan, Wang Heng, Huang Zhen. A New Method for Predicting Default Probability of Internal Rating Method in Commercial Banks: Research Based on Binary Response Panel Data Model [J] Southern Finance, 2011, (02): 21-28.

[4] Wei Hao, Zhao Tianxiang. Company Stock Market Crash Risk Warning Based on Text Mining: Empirical Evidence from Annual Reports' MD&A [J]. Journal of Fujian University of Business & Technology, 2022, (06): 33-41. DOI: 10.19473/j.cnki.1008-4940.2022.06.006.

[5] Zhang Jianwei. Inefficient Investment and Corporate Default Risk [J]. Investment Research, 2023, 42 (09): 130-144.

[6] Shen Long, Zhou Ying. Can management discussions and analyses predict Corporate default? -- Empirical Analysis Based on the Chinese Stock Market [J]. Journal of Systems Management, 2024, 33 (02): 441-459.

[7] Li Ran. Research on Multi-modal Financial Distress Prediction Based on Attention Mechanism [J]. Computer Programming Techniques and Maintenance, 2023, (11): 23-25 + 49. DOI: 10.16184/j.cnki.comprg.2023.11.024.

[8] Wu Shinan, Chen Zhiyu. Research on Bond Default Warning Model Based on Financial and Non-financial Information: Empirical Evidence from Machine Learning Methods [J]. Journal of Xiamen University (Philosophy and Social Sciences Edition), 2023, 73(06): 108-121.

DOI: 10.54254/2754-1169/2026.LD33165

[9] Hong Kanglong. Can BERT Artificial Intelligence Model Identify Opportunism in Management Tone? - Text Analysis Based on Listed Company Annual Reports [J]. Securities Market Herald, 2024, (10): 27-37 + 68.
[10] Wang Kemin, Wang Huajie, Li Dongdong, et al. Complexity of Annual Report Text Information and Managerial Self-interest - Evidence from Chinese Listed Companies [J]. Management World, 2018, 34(12): 120-132 + 194. DOI: 10.19744/j.cnki.11-1235/f.2018.0038.