The Emerging Role of Big Data and Machine Learning Technologies in Credit Scoring

Yuren Kong[1*]

[1] Gansu Hongyi Greenland Experimental High School, Lanzhou, China

Abstract. With the rapid development of big data and machine learning technologies, the application of sentiment analysis in financial credit scoring is gradually revolutionizing the traditional risk assessment paradigm. Traditional credit scoring models rely on structured financial data, making it difficult to capture the dynamic sentimental signals of borrowers and the potential impact of market sentiment. In this paper, this paper systematically explore the practical path of sentiment analysis technology to build a multi-dimensional credit risk portrait by mining unstructured data such as social media comments, news texts, customer feedback, etc., combined with machine learning algorithms (e.g., BERT, XGBoost, LSTM). Studies have shown that sentiment features can significantly improve the predictive ability of credit scoring models: in terms of technology integration, big data infrastructure (e.g., Hadoop, Spark) supports the real-time processing of massive text, while deep learning models (e.g., BLIP-NLP) realize the fine-grained extraction of sentiment signals. However, sentiment analysis in the financial domain still faces challenges such as data noise, dynamic adaptation, model interpretability, and privacy ethics.

1 Introduction

Credit scoring is a key tool for assessing the credit risk of borrowers in the modern financial services industry. Traditional credit scoring models mainly rely on historical financial data, personal credit records and socio-economic factors. Accompanied by advances in information technology and the rapid development of big data and sentiment analytics, financial institutions have begun to explore nontraditional data sources such as social media, consumer reviews, and news texts to optimize the credit assessment process for more nuanced and real-time information [1].

Sentiment analysis, an important application in the field of Natural Language Processing (NLP), aims to extract emotional tendencies (positive, negative, or neutral) from textual data, and machine learning and deep learning algorithms are typically employed to analyze and classify sentiment expressions. Consumer emotions and behaviors can profoundly affect their financial decisions, so the utilization of sentiment analysis techniques is particularly important for the financial sector. For example, Zhang et al. found that consumers' emotional responses to bank products, or changes in the market, may affect their demand for credit and ability to repay, thus indirectly affecting credit scores [2]. Sentiment analysis techniques are

* Corresponding author: vqfc7564@outlook.com

also able to analyze changes in sentiment in news reports, social media, and other unstructured data sources to reveal the potential impact of market sentiment on credit risk [3].

With the explosive growth of social media and online reviews, more and more financial institutions are utilizing these emerging “big data” sources for sentiment analysis to help build more comprehensive and accurate credit scoring models. By analyzing sentiment fluctuations on social media platforms, financial institutions can gain real-time insights into public perceptions of economic conditions or specific companies, and thus predict their credit risk [4]. In some cases, sentiment analysis has been used to monitor market sentiment and to dynamically adjust pricing strategies for financial products [5].

In addition to social media, sentiment analysis techniques have been employed to analyze customer reviews, product feedback, and consumer survey data that can provide additional credit information to banks and lending institutions [1]. Research has shown that sentiment analysis techniques can significantly improve the accuracy of credit scoring models by identifying the sentimental information implicit in the text [2]. Analyzing consumers' affective tendencies towards banking services can help banks to identify potentially risky customers and then take appropriate credit decisions [4].

Sentiment analysis techniques have been applied with many challenges. The expression of sentiment in textual data may vary depending on cultural background, language habits, and context, which increases the complexity of constructing sentiment analysis models [3]. Meanwhile, sentiment data in the financial domain is highly dynamic, and changes in sentiment may be affected by a variety of factors such as the economic situation, market fluctuations, and social events [2]. Therefore, designing a sentiment analysis model that can reflect the changes in market sentiment in real time is currently a hot topic in related research.

In this paper, will explore the potential of sentiment analysis technology in credit scoring by analyzing the existing research and application examples. The focus here will be on how to combine big data, machine learning techniques, sentiment analysis techniques and traditional credit scoring models to ultimately optimize financial decision-making and risk management strategies [5].

2 Integration of sentiment analysis and credit scoring models

Traditional credit scoring models mainly rely on structured financial data such as borrowers' income, debt levels, and credit history. While these data play an important role in assessing a borrower's repayment ability and credit risk, they have obvious limitations. First, structured financial data is usually historical and difficult to capture the borrower's current psychological state and behavioral patterns. Second, such data cannot reflect the immediate impact of market sentiment, social events, or changes in the macroeconomic environment on borrowers' credit risk. As a result, traditional models often lack real-time and dynamism in assessing credit risk, making it difficult to respond to the risks associated with rapid market changes [1,2].

With the popularization of big data technology, the financial industry has begun to integrate unstructured data such as social media comments, news reports and customer feedback. These unstructured data have the following characteristics: first, the data volume is large and comes from a wide range of sources, covering a variety of information in the borrower's daily life; second, the data are in various forms, including text, images, voice, etc.; third, the data are time-sensitive and can reflect the borrower's current psychological state and market sentiment. Mining the emotional signals in this unstructured data through sentiment analysis technology can effectively supplement the shortcomings of traditional models.

Sentiment analysis technology, through natural language processing (NLP) algorithms, can identify the emotional tendencies (positive, negative, or neutral) in the text and further reveal the borrower's psychological state and behavioral patterns [2]. For example, negative mood swings on social media may signal financial stress or psychological distress for borrowers, thereby increasing the risk of default [2]. Negative news reports may affect borrowers' confidence and spending behavior, which in turn affects their ability to repay. Dissatisfaction in customer feedback may imply that borrowers are dissatisfied with the financial institution's services or products, thereby reducing their willingness to repay [4].

The inherent logic of combining sentiment analysis techniques with traditional credit scoring models lies in the fact that sentiment data can provide dimensions of information that traditional financial data cannot. Specifically, sentiment analysis technology can supplement the shortcomings of traditional models in the following ways:

Real-time and dynamic: unstructured data can reflect the impact of market sentiment and social events on borrowers' psychological state in real time. For example, during periods of market turmoil, negative sentiment on social media may increase rapidly, and these changes can be captured promptly through sentiment analysis techniques and incorporated into credit scoring models, thus improving the accuracy of default risk prediction [6].

Mental States and Behavioral Patterns: Sentiment analysis not only identifies emotional tendencies in the text but also reveals the borrower's underlying mental states and behavioral patterns. For example, by analyzing mood swings on social media, the borrower's willingness and ability to repay can be predicted. This psychological level of analysis can provide a more comprehensive risk assessment for credit scoring models [2].

Multi-dimensional risk assessment: sentiment analysis techniques can integrate multiple unstructured data sources, such as social media comments, news reports, customer feedback, etc., to build a multi-dimensional risk portrait. This multi-dimensional risk assessment can reflect the borrower's credit risk in a more comprehensive way, not just based on historical financial data [1].

Enhance the adaptability of the model: Traditional credit scoring models are usually based on static financial data, which is difficult to adapt to the rapid changes in the market environment. Sentiment analysis techniques can make credit scoring models more adaptable and flexible by dynamically capturing sentiment signals, and better respond to market fluctuations and unexpected events [6,7].

Integrating unstructured data and mining sentiment signals through sentiment analysis technology can effectively complement the shortcomings of traditional credit scoring models. This combination not only improves the real-time and comprehensiveness of risk assessment but also provides financial institutions with more accurate and dynamic credit risk assessment tools.

Current research revolves around the optimization of sentiment analysis techniques and multimodal data fusion. Liu and Feng proposed to combine explainable artificial intelligence (XAI) with sentiment analysis to enhance the transparency and fairness of credit models by visualizing the decision weights of sentiment factors [8]. Mahmudul Hasan and Sachin Shetty developed the Lipschitz recurrent Neural Network (LRNN) and Generative Adversarial Network (GAN) hybrid framework, which solves the problem of insufficient model stability in traditional sentiment analysis and achieves accurate extraction of sentiment features in credit risk assessment [9]. In addition, the BLIP-NLP model was demonstrated to efficiently process social media and consumer feedback data, and its sentiment classification accuracy was significantly improved [10]. Nitin Rane et al. further validated the potential of deep learning techniques (e.g., BERT, LSTM) in real-time sentiment monitoring, and optimized the dynamics of the credit scoring by capturing the time-series correlation between borrowers' sentiment and repayment behavior adjustment mechanism [11].

These research results, while having their focus, collectively serve to improve the overall performance of credit scoring models. First, the application of XAI technology enhances the transparency and fairness of the model, enabling financial institutions to better understand

and explain the decision-making process of the model. Second, the hybrid framework of LRNN and GAN improves the stability and reliability of the model, laying the foundation for the wide application of sentiment analysis technology in the financial field. Again, the efficient data processing capability of BLIP-NLP model improves the quality of sentiment signals and provides richer data support for multimodal data fusion. Finally, the application of deep learning technology in real-time sentiment monitoring optimizes the dynamic adjustment mechanism of credit scoring, enabling the model to respond to market changes more flexibly.

These studies form a mutually complementary relationship in improving the transparency, stability, and accuracy of credit models. By integrating the results of these techniques, financial institutions can construct more comprehensive, dynamic, and accurate credit scoring models that can better respond to the complexity and uncertainty of the financial market. Future research should continue to deepen the integration and optimization of these technologies and further explore the potential of sentiment analysis technology in the financial field

The practical application of sentiment analysis in credit scoring still faces multiple complexity challenges. First, the subjective nature of sentiment expression significantly increases the difficulty of model parsing, and cultural differences, context-dependency, and ironic or ambiguous expressions commonly found in social media (e.g., surface positive comments implying negative emotions) may lead to biased sentiment classification [1]. Second, the problem of noise interference in unstructured data sources is prominent, and false information, emotional content and irrelevant text in social media need to rely on complex cleaning and feature screening techniques to ensure the reliability of sentiment signals [7]. On the technical level, the static architecture of traditional credit scoring models is difficult to adapt to the rapid changes in financial market sentiment, and although streaming computing frameworks (e.g., Flink) have been attempted to be used for real-time sentiment monitoring, there is still algorithmic lag in the deep fusion of historical data and dynamic sentiment features[9]. In addition, the interpretability flaws of deep learning models and privacy and ethical constraints form a double barrier: on the one hand, the opacity of the decision logic of black-box models hinders the regulatory compliance review, and the impact weight of sentiment factors on credit scoring needs to be revealed through interpretable AI (XAI) technology; on the other hand, sentiment data collection involves sensitive information such as users' social media behaviors, and under the data protection regulations such as the GDPR constraints, financial institutions need to balance data utility and privacy risk, which puts higher compliance requirements on model design [11]. Together, these challenges point to the key bottleneck for scaling sentiment analysis technology in financial scenarios.

3 Application of deep integration of big data and machine learning in credit scoring

The deep integration of sentiment analysis technology with big data and machine learning provides a breakthrough path for the innovation of credit scoring models. The traditional credit assessment system relies on structured financial data (e.g., repayment records, income levels), which makes it difficult to capture the emotional signals and dynamic risks in the borrower's behavioral patterns, whereas sentiment analysis significantly improves the granularity and forward-looking nature of risk assessment by mining unstructured text such as social media comments, consumption feedback, customer service records, etc., and combining with machine learning algorithms to construct a multi-dimensional risk portrait. Shahapur et al. classified 6,000 financial news articles into sentiment by random forest algorithm and improved the model accuracy to 82%, verifying the explanatory power of sentiment features on market fluctuations [6]; Khonde's team further demonstrated that combining the news headline sentiment index with the historical trading data can lead to 83% accuracy in predicting the market ups and downs, implying the potential association of sentiment factors on individual credit behavior [12]. The potential relevance of sentiment factors on individual credit behavior [12]. The synergistic effect of technology convergence is particularly significant: big data infrastructure (e.g., Hadoop, Spark) supports real-time processing of massive amounts of text, and Thakur builds a system that analyzes 210 million pieces of social media data per day [13]. Machine learning models achieve fine extraction of emotional signals. Jain et al. used the BERT algorithm to achieve 90.92% accuracy in financial text sentiment analysis, and its attention mechanism can accurately identify risky keywords such as "weakened willingness to repay", "tendency to spend impulsively", etc. [14]. Its attention mechanism can accurately identify risky keywords such as "weakened willingness to repay" and "tendency to consume impulsively" [14].

At the level of model construction, the current research shows an evolution from single data source analysis to multimodal fusion: Yu et al. combined XGBoost, LightGBM and TabNet models and used SMOTEENN technology to deal with data imbalance to improve the accuracy of credit risk prediction [15]. Kasture and Shirsath designed a hybrid RNN-LSTM framework to dynamically capture the risky keywords such as "willingness to repay" and "consumption impulse tendency" [16]. LSTM hybrid framework to dynamically capture the time-series correlation between borrowers' mood fluctuations and repayment behavior, which reduces the prediction error by 67% [16]. More cutting-edge practices are reflected in the construction of dynamic evaluation systems: based on the Flink stream processing engine, financial institutions can monitor changes in borrowers' social media sentiment (e.g., a surge in the concentration of negative sentiment) in real time, and trigger automatic adjustment of credit limits [17]. This dual-drive model of "static financial data + dynamic emotional trajectory" has been proven to have business value. In a consumer finance pilot project, the introduction of emotional stability coefficients increased the bad debt identification rate of high-risk groups by 11%, and the adjusted return of the risk pricing model increased by 22%.

However, the technology landing still faces multiple challenges: first, data ethics issues, such as emotional data collection needs to comply with privacy regulations such as GDPR [18]. Second, model interpretability requirements, deep learning black box characteristics may hinder regulatory review, need to introduce SHAP value and other methods to visualize the decision contribution of emotional factors; third, system compatibility bottlenecks, traditional financial institutions' batch processing architecture is difficult to support real-time Third, the system compatibility bottleneck, traditional financial institutions' batch processing architecture is difficult to support real-time sentiment analysis, and needs to be reconfigured to a Lambda architecture to balance offline and streaming computing. Nevertheless, sentiment analysis has shown the potential to reconfigure the logic of credit assessment by quantifying the "emotion-behavior-risk" chain, promoting credit scoring from "static financial observation"to"dynamic psychological insight", and providing a new way for financial risks to be assessed. By quantifying the "emotion-behavior-risk" chain, it can shift credit scoring from"static financial observation"to"dynamic psychological insight", providing a new paradigm for the accurate prevention and control of financial risks

In general, different studies and practices have focused on improving the accuracy of credit risk prediction and reducing the prediction error etc. Shahapur et al.'s study focuses on the analysis of market sentiment, while Khonde's team further explores the correlation between the sentiment factor and individual credit behaviors. Yu et al.'s multimodal data fusion method is excellent in terms of prediction accuracy, while the dynamic analysis framework of Kasture and Shirsath is excellent in terms of prediction accuracy. Shirsath's dynamic analysis framework is more advantageous in handling time series data. A real-time monitoring practice based on Flink demonstrates the great potential of sentiment analysis techniques in real business. Together, these research and practice cases promote the application and development of sentiment analysis technology in the credit scoring field and provide rich references for future research and practice.

4 Conclusion

This paper systematically explores the application of sentiment analysis techniques in financial credit scoring and their deep integration with big data and machine learning techniques. It is found that sentiment analysis can significantly improve the predictive ability of credit scoring models by mining sentiment signals in unstructured data such as social media comments, news texts and customer feedback, combined with machine learning algorithms (e.g., BERT, XGBoost, LSTM). Big data infrastructures (e.g., Hadoop, Spark) provide support for real-time processing of massive text, while deep learning models (e.g., BLIP-NLP) enable fine-grained extraction of sentiment signals. However, sentiment analysis still faces many challenges in the financial domain, including data noise, dynamicity adaptation, model interpretability and privacy ethics. Nonetheless, sentiment analysis has demonstrated the potential to reconfigure the logic of credit assessment and promote the paradigm shift of credit scoring from “static financial observation”to “dynamic psychological insight”.

Given the current application of sentiment analysis in financial credit scoring, this paper suggests that future research should focus on multimodal data fusion, optimization of real-time sentiment monitoring architecture, and deepening the application of interpretable AI techniques. Multimodal data fusion can integrate text, image, voice and other data sources to further enrich the dimensions of sentiment analysis; the optimization of real-time sentiment monitoring architecture can improve the response speed of the model to changes in market sentiment; and the application of interpretable AI technology can help solve the black-box problem of deep-learning models, and enhance the transparency and credibility of the models. In addition, with the continuous development of technology, financial institutions should actively explore the application of sentiment analysis technology in other financial business areas, such as risk early warning and investment decision-making, to give full play to its value in the financial field. In the future, sentiment analysis technology is expected to realize wider application and deeper integration in the financial field, and provide more powerful support for the accurate prevention and control of financial risks.

References

1. S. Chakraborty, Sentiment analysis in the perspective of natural language processing. Int. J. Res. Appl. Sci. Eng. Technol. (2023)
2. J. Zhang, An overview of the application of sentiment analysis in mental well-being. Appl. Comput. Eng. (2023)
3. Z. Li, T.-E. Lin, Y. Wu, M. Liu, F. Tang, M. Zhao, Y. Li, Unified generative framework for sentiment analysis. UniSA (2023)
4. P. Agarwal, P. Shrivastava, Twitter sentiment analysis using machine learning. Computology (2021)
5. A.K. Modi, Sentiment analysis on textual data - A comparison of accuracy using different algorithms. Indian Sci. J. Res. Eng. Manag. (2024)
6. S.S. Shahapur, A. Koralli, G. Chippalakatti, M.M. Balikai, D. Mudalagi, R. Dias, S. Devali, K. Wajantari, Discovering untapped potential in finance markets using NLP-driven sentiment analysis. Indian J. Sci. Technol. (2024)

7. J. Wang, W. Zhao, F. Meng, G. Qu, Sentiment analysis methods based on multi-modal fusion and deep learning. 13th IEEE Int. Conf. Commun. Syst. Netw. Technol. (2024)
8. D. Liu, F. Feng, Integrating explainable AI for fair and transparent financial decision-making. Adv. Credit Scoring Models (2024)
9. M. Hasan, S. Shetty, Sentiment analysis with Lipschitz recurrent neural networks based generative adversarial networks. 2024 Workshop Comput. Netw. Commun. (2024)
10. P.V.V.S. Srinivas et al, BLIP-NLP model for sentiment analysis. Proc. Second Int. Conf. Edge Comput. Appl. (2023)
11. N. Rane, S. Choudhary, Artificial intelligence, machine learning, and deep learning for sentiment analysis in business to enhance customer experience, loyalty, and satisfaction. J. Indones. Manaj. Inform. Komun. (2024)
12. S.R. Khond, S.S. Virnodkar, S.B. Nemade, M.A. Dudhedia, B. Kanawade, S.H. Gawande, Sentiment analysis and stock data prediction using financial news headlines approach. Rev. Intell. Artif. (2024)
13. S.N. Thakur, S. Maurya, M.K. Singh, M. Koranga, Concepts, approaches, and challenges in a study of sentiment analysis. 15th IEEE Int. Conf. Comput. Intell. Commun. Netw. (2023)
14. A. Jain, S.S. Bisht, S. Arora, B. Anand, A comparative study on deep learning techniques using sentiment analysis. 2023 5th Int. Conf. Adv. Comput. Commun. Control Netw. (2023)
15. C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, S. Meng, Advanced user credit risk prediction model using LightGBM, XGBoost and TabNet with SMOTEENN. arXiv (2024)
16. P. Kasture, K. Shirsath, Enhancing stock market prediction: A hybrid RNN-LSTM framework with sentiment analysis. Indian J. Sci. Technol. (2024)
17. R. Patidar, S. Patel, Sentiment analysis using Hadoop framework and machine learning approach. 2023 Third Int. Conf. Ubiquitous Comput. Intell. Inf. Syst. (2023)
18. J.S. Hutagalung, R. Rasiban, Analisis sentimen keuangan (Data FiQA and Financial PhraseBank) menggunakan algoritma logistic regression dan support vector machine. J. Indones. Manaj. Inform. Komun. (2023)