Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 4, No. 2, 2025

Research on the Application of Multimodal Information Fusion in Financial Data Prediction

Qingying Zhi

School of Computing and Data Science, The University Of Hong Kong, Hong Kong, China
u3665385@connect.hku.hk

Abstract: The high complexity and uncertainty of financial markets make it difficult for predictive models that rely on a single data source to effectively address these challenges. Multimodal information fusion, by integrating heterogeneous data sources such as text, numerical data, images, and time series, reveals hidden connections and opens new avenues for improving the accuracy and robustness of financial forecasts. This paper systematically analyzes the theoretical foundations, key technical approaches, and typical scenarios for applying multimodal information fusion to financial forecasting. The study first analyzes the characteristics and categories of multimodal financial data, encompassing structured market data, unstructured text, time series data streams, and visual information. Secondly, it focuses on the core technologies of multimodal fusion, covering the data layer, feature layer, decision layer, and hybrid fusion strategies. It also analyzes the advantages of deep learning models in processing and fusing heterogeneous data. Through specific application cases such as stock price trend forecasting, credit risk assessment, and macroeconomic indicator forecasting, this paper demonstrates the significant effectiveness of multimodal fusion in capturing market sentiment, identifying potential risks, and improving forecasting accuracy. The study also identifies key challenges currently faced, including data heterogeneity, model interpretability, computational efficiency, and privacy protection. Finally, we outline future research directions, emphasizing the importance of cross-modal alignment, adaptive fusion mechanisms, interpretability enhancement, and privacy-preserving fusion frameworks. This study provides theoretical support and practical references for deepening the application of multimodal fusion in intelligent financial decision-making.

Keywords: Multimodal information fusion, financial forecasting, deep learning, heterogeneous data, market sentiment analysis, risk modeling, explainable artificial intelligence.

1. Introduction

The core of financial market forecasting lies in accurately assessing asset prices, risk levels, and economic trends. Traditional models primarily rely on structured time-series data such as historical prices and trading volumes. However, as complex systems, financial markets are driven by a vast amount of heterogeneous information, far beyond the scope of traditional quantitative indicators. Multiple sources of information, including news announcements, policy documents, social media sentiment, company financial reports, analyst opinions, and even satellite imagery, all contain key market signals. While these data may vary in form, they are inherently interconnected. For example, breaking negative financial news often quickly sparks heated discussion on social media platforms, ultimately leading to unusual stock price fluctuations.

Single-modal models often struggle to effectively capture complex cross-modal connections and dynamic interactions. Quantitative analysis excels at identifying patterns in numerical sequences but is less sensitive to the emotional fluctuations and event shocks inherent in text. While text parsing can interpret semantic meaning, it struggles to accurately quantify the time lag and intensity of its impact on market fluctuations. This fragmentation leads to insufficient model information utilization, significantly reducing predictive capabilities in the face of unexpected events or emotional market conditions.

Multimodal fusion offers a new approach to addressing this bottleneck. The key lies in breaking down data silos and leveraging machine learning to collaboratively integrate financial data from diverse sources and in various formats, extracting complementary information and generating more comprehensive market insights. Recent advances in natural language processing, computer vision, and deep learning fusion architectures have significantly boosted the application of these technologies in financial forecasting. Practice has demonstrated that predictive models that integrate news sentiment and historical prices, or risk control models that combine financial data and related-party public opinion, outperform single-modality models.

This study systematically reviews the research progress of multimodal information fusion technology in the field of financial data forecasting. By analyzing data characteristics, exploring core fusion technologies and model architectures, evaluating application results, and analyzing existing challenges and future directions, it aims to provide a reference for building more intelligent and robust forecasting systems, helping FinTech move towards a new stage driven by data fusion.

2. Characteristics and Types of Multimodal Data in the Financial Sector

Financial markets naturally harbor diverse data modalities. Primarily, structured numerical and time series data encompasses the prices, trading volumes, order book depth, order flow, and various technical and macroeconomic indicators for assets such as stocks and bonds. This highly structured data, primarily consisting of time-aligned numerical sequences, is the cornerstone of quantitative analysis, and its time-dependence and volatility are key to modeling.

Secondly, there is a vast volume of unstructured text data. This primarily includes: real-time updates from financial news agencies covering company developments, policy changes, and geopolitical events, which are highly timely; investor discussions on social media reflecting market sentiment and buzz; regular reports and announcements from listed companies disclosing financial and operating details; analyst research reports providing professional interpretations and forecasts; and legal and regulatory documents issued by regulatory agencies. These texts contain rich semantics, sentiment, event types, and entity relationships, which can be converted into structured predictive features through natural language processing.

The third important modality is visual data. For example, financial charts inherently contain information such as price patterns, support and resistance levels, and computer vision can be used to identify these patterns. Remote sensing data such as satellite imagery and aerial photography monitor economic activity and serve as alternative data for predicting company performance or product supply. The visual and vocal characteristics of speakers in company press conference videos provide an additional dimension for capturing market sentiment.

At the same time, the importance of network data continues to grow [1]. This encompasses: equity, supply chain, and competitive and cooperative networks among companies; investor social circles, attention chains, and capital flow networks; and counterparty risk networks among financial institutions. This type of graph-structured information can clearly reveal risk transmission chains, information diffusion pathways, and systemic market connections. Graph neural networks (GNNs) are an effective tool for processing these data. These four types of data, each with their own unique characteristics, complement each other and jointly support the integration of multimodal financial information, creating a three-dimensional picture of the market.

3. Core Technologies and Models for Multimodal Information Fusion

Transforming heterogeneous financial multimodal data into predictive capabilities lies in the selection of fusion technology and model architecture. Depending on the stage of fusion, the main technical approaches include:

Data-layer fusion (early fusion): Directly integrating data at the raw or low-level feature level. For example, after aligning timestamps, vectorized news text is concatenated with standardized time series data such as stock prices and trading volume to form a unified feature vector, which is then fed into the prediction model. This approach offers the advantage of enabling the model to learn cross-modal correlations directly from raw data interactions, minimizing information loss. However, the main challenge lies in the significant differences in scale, frequency, and dimensionality between modalities. Direct concatenation can easily lead to information flooding or the curse of dimensionality, and strict temporal alignment requirements are difficult to implement.

Feature-layer fusion (mid-term fusion): The current mainstream strategy. Each modal data is first subjected to a dedicated encoder to extract high-level features: time series data is often encoded using RNNs, TCNs, or temporal transformers; text uses CNNs, RNNs, or pre-trained language models; images use CNNs or visual transformers; and graph data uses GNNs. These features are then fused using methods such as concatenation, weighted averaging, dynamic weighting based on attention mechanisms, tensor operations, or dedicated fusion networks. This strategy allows each modal feature to fully learn its own characteristics before capturing inter-modal correlations through fusion, offering both flexibility and robustness. Decision-level fusion (late fusion): Prediction models are trained independently for each modality, each generating its own predictions, which are then fused at the decision level. Fusion methods include voting, weighted averaging, stacking, or Bayesian probabilistic fusion [2]. This approach is relatively simple to implement, easily integrating with existing single-modal models, and requiring low inter-modal independence. However, its primary drawback is that the models are trained independently, failing to effectively leverage the potential complementary information between modalities for joint optimization, potentially missing important correlations.

Hybrid fusion strategy: This strategy combines the aforementioned strategies to leverage their respective strengths. For example, closely related modalities can be fused at the feature level, and then other modalities can be fused at the decision level. Deep learning models are a key support mechanism, particularly the multimodal Transformer architecture. Its self-attention mechanism is naturally well-suited to sequential data, and its cross-modal attention layer explicitly models the interactions between features from different modalities. Graph neural networks (GNNs) excel at processing relational networks or constructing multimodal graphs containing entity relationships. Multi-task learning frameworks facilitate the extraction of universal multimodal features at the shared layer by jointly optimizing related tasks. The choice of technology and architecture should be closely integrated with the specific prediction task requirements, data modality characteristics, and data quality.

4. Typical Application Scenarios of Multimodal Fusion in Financial Forecasting

Multimodal information fusion technology significantly improves model performance in several core areas of financial forecasting by integrating different types of data:

Stock Price and Market Index Forecasting: Traditional methods rely on historical prices and trading volume, but short-term fluctuations are often driven by sentiment and events. Fusion technology integrates data such as social media text, company fundamentals, search trends, and option volatility. For example, models that combine analysis of historical price series with text sentiment can more accurately capture the dynamic impact of sentiment on prices, outperforming single-price models in predicting short-term trends and volatility, especially during periods of market volatility.

Credit Risk Assessment: Traditional models primarily rely on financial data and credit records. Multimodal fusion incorporates corporate announcements, negative events reported in news reports, negative social media comments about the company or its supply chain, and inter-company networks. For those with limited credit records, online business data can also be integrated as a supplement. This not only enables earlier identification of risks at large companies, but also provides a more comprehensive perspective for assessing clients with limited credit records.

Macroeconomic Indicator Forecasting: Forecasting key

indicators such as GDP and CPI is crucial for decision-making. Traditional models rely on statistical data released with a lag. Fusion technologies incorporate real-time or near-real-time information: analyzing news and policy texts, satellite imagery, social media data, and financial market data reflecting market expectations. Models built from this data can provide more timely and, in some cases, more accurate predictions of macro trends [3].

Trading Strategies and Risk Management: In quantitative trading, integrating technical price indicators, public sentiment signals, and market implied volatility generates smarter trading signals. In risk management, integrating market risk models, counterparty public sentiment risk signals, and financial network structure information can build a more comprehensive and dynamic risk measurement and early warning system. Furthermore, integrating trading patterns, user behavior, device information, and customer service texts enhances fraud detection capabilities and identifies complex tactics.

5. Challenges and Limitations

Despite its promising prospects, multimodal information fusion faces a series of significant challenges in its practical application in financial forecasting:

Data-Level Challenges: The heterogeneity of financial multimodal data is a fundamental challenge. Data from different modalities vary significantly in acquisition frequency, data structure, and semantic expression, making effective alignment and fusion extremely difficult. Data sparsity and noise are prominent issues, especially since high-quality alternative data is expensive to acquire and lacks comprehensive coverage. Social media content contains a large amount of irrelevant information, noise, and intentionally misleading content. News data may be subject to reporting delays or bias. Data timeliness is crucial, and given the rapid response of financial markets, accurately matching news release times, social media buzz, and market transaction times is crucial. Furthermore, the storage, processing, and computational costs of massive amounts of multimodal data are high, posing challenges to infrastructure.

Model-Level Challenges: Current mainstream deep fusion models are often viewed as "black boxes," making their internal decision-making logic difficult to clearly explain. In the highly regulated financial sector, where clear accountability is emphasized, the lack of model interpretability seriously hinders business implementation and regulatory compliance. Regulators and risk management personnel need to understand why models make specific predictions, particularly how multimodal features interact. Model complexity coexists with the risk of overfitting. Deep fusion models have numerous parameters. When data is limited or incompletely representative of the population, they are prone to overfitting to noise or specific patterns in the training set, resulting in reduced generalization to real-world market conditions or out-of-sample data. Robustness to missing or invalid modalities is insufficient. In real-world applications, temporary loss of data from certain modalities or signal failure of specific modalities often occur. How the model adaptively adjusts fusion weights or relies solely on available modalities to make reliable predictions remains an unresolved issue [4].

Challenges at the application and ethical levels: The primary difficulty is identifying the value of fused information. Not all integrated data has actual predictive power. The key lies in designing mechanisms to automatically screen core modalities and features, avoid interference from redundant information, and focus on truly effective signals. Data privacy and security issues are particularly prominent. Integrating multi-source data faces stringent privacy regulations. Ensuring compliance, effective anonymization, and data security during the integration process are key challenges. Regulatory compliance and audit thresholds are high. The strong regulatory nature of the financial industry requires complex fusion models to meet auditability, traceability, and fairness requirements, significantly increasing the complexity of design and deployment. Furthermore, the convergence of model strategies can trigger systemic risks. If market participants widely adopt similar multimodal fusion methods, it can easily trigger a herd effect and exacerbate market volatility.

6. Solutions and Future Research Directions

Improving data quality and fusion robustness: Solving cross-modal semantic alignment is key. Leveraging knowledge graphs to construct financial entities, events, and their relationships provides a unified semantic anchor for data from different modalities, enabling precise association. Furthermore, fusion architectures robust to noise and missing modalities can be developed, for example by introducing gating mechanisms to dynamically adjust the weights of each modality or applying generative models to reasonably estimate missing data. Continuously optimize key aspects of sentiment analysis in financial text, such as accuracy and domain adaptability.

Improving model interpretability and trustworthiness: This is crucial for the implementation of financial applications. Prioritizing the application of Explainable Artificial Intelligence (XAI) technologies: Utilizing attention mechanism visualization and feature importance analysis to reveal the specific modalities and features that model decisions rely on; exploring the design of interpretable fusion mechanisms, such as modularization and rule-based guidance. Developing causal inference frameworks aims to distinguish correlation from causation, clarify the actual transmission path of information impact, and enhance model stability and decision credibility. Promoting the application of model transparency standards and audit tools that meet financial regulatory requirements.

Exploring New Fusion Architectures and Learning Paradigms: Model architecture innovation is a key breakthrough direction. The focus is on developing adaptive fusion frameworks that enable models to dynamically adjust fusion strategies and modal weightings based on data characteristics or task context. Deep fusion of graph neural networks (GNNs) and Transformers has become a research hotspot: embedding multimodal data into a unified graph structure, using GNNs to aggregate local neighborhood information, and Transformers to capture global long-range dependencies, effectively modeling complex networks of connections between financial entities. Federated learning offers a new approach to addressing data silos and privacy concerns, allowing multiple parties to collaboratively train fusion models by exchanging encrypted gradients or model parameters while retaining the original data locally. Meta-learning and continuous learning techniques aim to enhance models' ability to rapidly adapt to new market environments

(such as bull-bear transitions) and the emergence of new data types.

Expanding Applications and Deepening Governance: Future research will focus on exploring the in-depth application of multimodal fusion in complex scenarios such as high-frequency trading, algorithmic market making, and portfolio optimization. This technology holds great potential in the field of financial regulatory technology (RegTech) [5]. For example, it can integrate trading patterns, communication records, and network relationships to identify abnormal behaviors such as market manipulation in real time. At the same time, a supporting ethical and governance framework is urgently needed: formulate ethical guidelines to regulate data collection and use; develop and embed effective privacy protection technologies; systematically assess and mitigate potential model biases to ensure fair decision-making; and promote the implementation of adaptable regulatory sandboxes and risk assessment standards to strike a balance between stimulating innovation and controlling systemic risks. Addressing these complex challenges requires collaborative advancement across multiple fields, including finance, law, and ethics.

7. Conclusion

Multimodal information fusion has become a key path to improving financial forecasting capabilities. By integrating diverse and heterogeneous data such as structured market trends, unstructured text, visual information, and relationship networks, it breaks through the limitations of a single source and builds a more comprehensive and dynamic market cognition system. Empirical research has shown that deep learning-based fusion solutions significantly surpass traditional single-modal models in terms of forecast accuracy and robustness in core scenarios such as stock price forecasting, credit risk assessment, and macroeconomic outlook. They can effectively capture market sentiment and potential risk factors, improving the timeliness and accuracy of forecasts.

However, the practical application of this technology still faces significant challenges. Data heterogeneity, noise, data sparsity, and strict time-series alignment requirements pose fundamental obstacles. At the model level, the lack of interpretability due to its "black box" nature hinders trust building in a highly regulated financial environment. Furthermore, the risk of overfitting and sensitivity to modality loss arising from excessive model complexity cannot be ignored. Applications must also address practical issues such as privacy compliance, data security, effective screening of key information, and potential ethical and fairness considerations. These interrelated challenges constitute key bottlenecks for practical application.

Future research should focus on developing more robust and adaptive cross-modal semantic alignment mechanisms; deepening the application of Explainable Artificial Intelligence (XAI) technologies to enhance model transparency and decision credibility; exploring federated learning architectures to facilitate cross-institutional data collaboration while strictly protecting privacy; and developing dynamic meta-learning and continuous learning model mechanisms to effectively respond to market changes and the introduction of new modal data. Furthermore, a comprehensive ethical framework and governance framework must be established to ensure that this technology, while improving predictive performance, proactively meets regulatory compliance requirements, effectively protects the rights and interests of data subjects, and maintains market fairness and stability. With the continuous iteration of technology and the deepening of interdisciplinary collaboration, multimodal information fusion is expected to become the core engine driving the next generation of intelligent financial decision-making systems, injecting continuous impetus into refined risk management, investment strategy optimization and overall market efficiency improvement.

References

[1] Tang Li. Information fusion and computational intelligence model for financial time series prediction [D]. University of Electronic Science and Technology of China, 2018.

[2] Wang Min. A preliminary study on dynamic information fusion method for financial risk early warning system [J]. Financial Economics, 1999, (05): 28-29.

[3] Gao Haiyan. Dynamic information fusion method for financial risk early warning system [J]. Computer and Information Technology, 1999, (01): 60-61. DOI: 10.19414/j.cn ki.1005-1228.1999.01.020.

[4] Tang Li. Information fusion and computational intelligence model for financial time series prediction [D]. University of Electronic Science and Technology of China, 2018.

[5] Gao Haiyan. Dynamic information fusion method for financial risk early warning system [J]. Computer and Information Technology, 1999, (01): 60-61. DOI: 10.19414/j.cnki.1005-1228.1999.01.020.