Transactions on Computational and Scientific Methods | Vo. 4, No. 6, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Forecasting Asset Returns with Structured Text Factors and Dynamic Time Windows

Xin Su
University of Chicago, Chicago, USA
xinsuxs@outlook.com

Abstract: This study addresses the limitations in asset allocation modeling related to insufficient use of unstructured information and the absence of dynamic modeling mechanisms. It proposes a regression algorithm for asset returns that integrates macro text factors with time window modeling. The method takes macroeconomic texts as input and extracts high-dimensional semantic vectors using a pre-trained semantic encoder. These vectors capture key signals such as policy direction, market expectations, and structural changes in the economy. During factor representation, a residual fusion mechanism is introduced to enhance the nonlinear expressiveness of semantic features. A unified time window structure is then constructed to model the dynamic influence of text signals on asset returns over continuous periods. To improve prediction accuracy and responsiveness, a nonlinear regression layer is applied after sequence aggregation, enabling continuous forecasting of asset returns. This paper designs multiple experiments on a public macro text dataset and various types of asset return data. Systematic sensitivity analyses are conducted across several dimensions, including encoding size, asset class, text noise intensity, and corpus source heterogeneity. The experimental results show that the proposed model outperforms several mainstream baseline methods in regression accuracy, stability, and cross-market adaptability. The results confirm the effectiveness of combining macro text factors with temporal structures for asset return prediction and demonstrate strong structure-aware modeling performance.

Keywords: Text Regression Modeling, Semantic vector representation, Time window mechanism, Asset Allocation Forecast

1. Introduction

Asset allocation has long been a core issue in financial decision-making, attracting significant attention from both academic and practical domains. Traditional asset allocation models often rely on economic theory, such as the mean-variance model, the Black-Litterman model, and dynamic optimization approaches. These models typically assume stable market conditions and complete information structures[1]. However, with the increasing complexity of the macroeconomic environment and the rising volatility in financial markets, static allocation methods are proving inadequate in handling nonlinear fluctuations, information lags, and differences in investor expectations. As the cost of information acquisition decreases, the availability of large-scale financial text data offers new possibilities for modeling investor expectations and market trends. Financial texts contain rich signals of market sentiment, macroeconomic information, and structural implications. Such unstructured data are becoming important sources of factors that influence asset returns[2].

In recent years, the rapid development of natural language processing techniques has provided strong support for modeling financial texts. Attention has shifted from traditional numerical factors to textual sources such

as news articles, company announcements, policy statements, and analytical reports. In particular, macroeconomic texts, including central bank statements, fiscal outlooks, and macroeconomic research papers, affect asset price expectations and allocation behavior in multiple ways. These texts offer interpretations of economic cycles, risk preferences, and interest rate trends, and they often carry forward-looking information and sentiment signals. However, extracting predictive factors from massive macro texts and incorporating them into asset allocation strategies suitable for dynamic markets remains a major challenge in financial modeling[4].

In asset allocation modeling, the design of time windows plays a critical role in capturing market dynamics and enhancing model adaptability. Financial markets respond to information with delay and persistence. Investors differ in how quickly and strongly they react to new information. As a result, the impact of macro signals on returns often exhibits multi-scale and multi-phase patterns. Static regression frameworks are insufficient to describe these dynamics. It is essential to introduce time-series-based modeling mechanisms to capture the temporal variation of factor influence. The design of time windows affects both the explanatory power of text-based factors and the model's robustness and generalization across market cycles. Proper time structures improve prediction accuracy and interpretability, enabling asset allocation models to respond dynamically to market changes[5].

Moreover, macro text factors are highly heterogeneous, high-dimensional, and semantically complex. Traditional factor construction methods struggle to extract their latent predictive signals. Deep semantic modeling, such as using pre-trained language models for text encoding, can capture nuanced expressions of policy statements, economic expectations, and financial opinions across multiple semantic layers. This enhances the model's ability to identify macro risks and opportunities. At the same time, jointly modeling text information and asset price time series in a regression framework allows the model to integrate structured market data with unstructured text signals. This fusion improves the effectiveness and responsiveness of asset allocation strategies. By combining semantic factors with temporal modeling, it becomes possible to build more accurate models of asset returns.

In today's financial environment, marked by high-frequency volatility and frequent policy shocks, the robustness and adaptability of asset allocation models are increasingly important. Text data serves as a crucial bridge to understanding policy directions and market expectations. They offer new perspectives and modeling dimensions for asset allocation. Research on asset return regression algorithms based on macro text factors and time window modeling represents both an extension of current financial modeling paradigms and a response to the evolution of financial technology. This line of research promotes a shift in asset allocation from static to dynamic and from numerical to semantic approaches. It also lays a theoretical and technical foundation for building more resilient and forward-looking financial decision systems.

2. Related Work

In recent years, asset allocation modeling has gradually evolved from traditional statistical paradigms toward approaches that integrate data-driven techniques and semantic modeling. Classical asset allocation methods are mainly built on mean-variance theory, capital asset pricing models, or dynamic programming frameworks. These methods emphasize optimization based on structured financial data such as historical returns, volatility, and covariance matrices. However, they often assume market stability and rational investor behavior, which limits their ability to respond to complex policy shocks and sentiment transmission in real-world markets. Under high macroeconomic uncertainty, traditional models are slow to react to sudden events and information shifts. They fail to capture latent risk signals and structural changes embedded in textual data. As a result, more studies are introducing financial texts into the asset allocation process to improve prediction performance and strategy robustness[6].

The rise of text-based factor modeling has significantly broadened the landscape of financial factor research by introducing unstructured textual information as a valuable signal source. Early efforts in this domain primarily focused on simple techniques such as word frequency statistics, sentiment lexicons, or sets of

keywords manually defined by domain experts. These methods were commonly used to extract sentiment or event-driven indicators from financial texts, including news articles, corporate announcements, and research reports. While effective to a limited extent, such approaches could not often capture deeper semantic content or reflect the temporal dynamics inherent in financial discourse[7].

As natural language processing technologies have evolved, more advanced methods have emerged to enhance the semantic richness and time-aware representation of financial texts. Techniques such as word embeddings, topic modeling, and syntactic analysis have been applied to construct more nuanced textual features. More recently, high-dimensional outputs from pre-trained language models have been directly integrated into regression frameworks, offering greater flexibility and generalization than static text factors. This is particularly valuable when dealing with macroeconomic texts, which tend to be lengthy, abstract, and indirectly phrased. Such characteristics challenge traditional linear factor modeling, making it essential to adopt more refined semantic encoding strategies to accurately extract the latent financial insights embedded in the text.

To capture the dynamic relationship between financial texts and asset prices, some studies have introduced time series modeling to account for lagged factor effects. Mainstream methods include sliding window regression, dynamic factor models, varying coefficient regressions, and recurrent neural networks. These techniques aim to describe how text signals influence asset returns across different time scales and temporal patterns. Several studies point out that financial markets tend to respond to macro policies or sentiment signals with delay and in stages. Static factor construction at a single time point often introduces distortion risks. Introducing a time window mechanism improves factor robustness and reveals the linkage between policy expectations and market adjustments over time. The current trend is to integrate text modeling with temporal modeling to improve both structural expressiveness and generalization in asset allocation models[8].

Despite progress in text factor modeling and time series prediction, most existing methods still face limitations such as shallow semantic understanding, weak temporal modeling, or unstable factor structures. In practice, many studies continue to use fixed windows or static factor strategies, which fail to capture the nonlinear market responses to policy texts. Some models also lack sufficient selectivity and interpretability when dealing with redundant texts, semantic ambiguity, or co-occurring events. Thus, building a regression framework that can jointly model text semantics and temporal structures, while enabling dynamic factor representation and continuous prediction, remains a key challenge in asset allocation modeling. This requires deeper and more accurate semantic extraction, as well as more flexible temporal mechanisms. Together, they can enhance the model's adaptability to complex market conditions and improve the precision of return prediction.

3. Model Design

This study proposes an asset return rate regression algorithm that integrates macro text factors with time window modeling, aiming to capture the dynamic influence of unstructured macroeconomic information on financial markets. The overall method is composed of three core components. First, a semantic encoding module extracts high-dimensional representations from macroeconomic texts using a pre-trained language model, capturing key signals such as policy direction, economic outlook, and market sentiment. Second, a structured time window mechanism organizes the encoded text features into sequential input segments, reflecting the temporal dynamics and delayed effects of macro information. Third, a regression mapping module aggregates the temporally structured features and transforms them into continuous asset return predictions through a nonlinear function. The full architecture of the proposed model is illustrated in Figure 1.

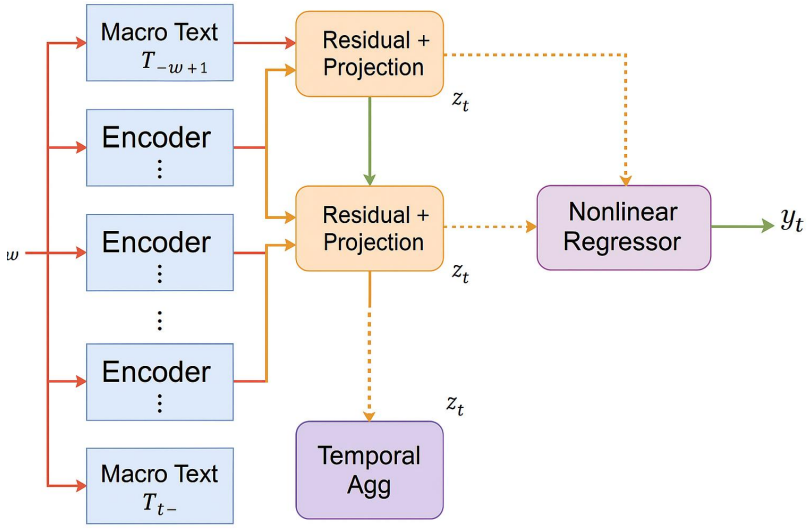

Figure 1. Framework of the Macro Text-Driven Dynamic Allocation Predictor

First, for each macro text T_t at time t, use the pre-trained language model to extract its high-dimensional semantic representation $h_t \in R^d$, that is:

$$h_t = Encoder(T_t)$$

Encoder represents a language encoder with fixed parameters, such as the Transformer structure, which can perform deep semantic modeling of economic expectations and policy intentions in macro narratives. In order to enhance the model's selectivity for key information in long texts, a residual fusion mechanism and nonlinear projection are introduced to construct the final semantic vector.

$$z_t = RELU(W_1 ht + b_1) + h_t$$

Where W_1 and b_1 are trainable parameters, and $RELU$ is the activation function.

Considering the time-dependent characteristics of asset returns, this method introduces a time window mechanism when building a regression model to perform sequence modeling on the text vectors at historical moments w. Specifically, a window sequence $\{z_{t-w+1},...,z_t\}$ is constructed, and a time-sensitive aggregate representation is generated through a time series modeler (such as a gating structure or an attention module):

$$s_t = TemporalAggg(z_{t-w+1},...,z_t)$$

This representation captures the evolution of text signals in a continuous time interval, which helps to describe the dynamic response relationship between the market and macro text. At the same time, in order to improve the contextual consistency of the representation, a multi-head attention mechanism is used to enhance the temporal modeling process.

In the regression prediction stage, the model introduces a nonlinear mapping layer to map the sequence representation into continuous return rate prediction values, as follows:

$$\hat{y}_t = w_r^T \cdot \tanh(W_2 s_t + b_2)$$

w_r、W_2、b_2 is the training parameter and $\tanh$ provides nonlinear transformation capability. Regression training uses minimizing the mean square error as the optimization goal, and the loss function is defined as:

$$L = \frac{1}{N}\sum_{t-1}^{N}(y_t - \hat{y}_t)^2$$

Where y_t is the actual return rate and $\hat{y}_t$ is the model prediction value. This loss function is more sensitive to large errors, which helps to suppress extreme prediction deviations and improve overall stability.

In summary, this method constructs a structure-aware regression prediction model by combining text semantic modeling with a time window mechanism, which can dynamically extract market-driven signals from macro texts and map them to the continuous value domain of asset returns. The entire model framework supports end-to-end training and can be adapted to a variety of text sources and asset categories, with strong flexibility and scalability.

4. Experimental Data Preparation

This study uses the Global Macro Forecast Text Dataset (GMFTD) as the source of macroeconomic textual data. The dataset consists of various types of macroeconomic texts, including central bank statements, government fiscal policies, macro research briefs, and official economic outlooks. It features a stable structure, broad coverage, and a long period. The texts are arranged in chronological order and aligned with daily return data of publicly traded assets, making it naturally suited for time series forecasting tasks.

The texts in the GMFTD mainly originate from official releases of G20 major economies and reports by international organizations. All content is in English and covers macroeconomic events and policy changes from 2005 to the present. Each text is accompanied by an accurate timestamp, which allows precise alignment with financial market data. The texts also contain a strong semantic structure, including policy orientation, economic expectations, and market sentiment. These features help the model capture underlying driving mechanisms.

To support the regression modeling task, this study synchronizes the texts in GMFTD with standard asset return data. Representative asset classes such as stock indices, sovereign bonds, and gold prices are selected as target variables for prediction. All texts are preprocessed before input. This includes sentence segmentation, noise removal, encoding, and length truncation, ensuring semantic consistency and feasibility for temporal modeling. The dataset has been widely used in studies related to financial text mining and macroeconomic expectation modeling. It offers strong representativeness and empirical value.

5. Performance Results and Discussion

In the experimental results section, the relevant results of the comparative test are first given, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	**MSE**	**R2**	**IR**
Portfolio Transformer[9]	0.0253	0.472	0.47
Text-Based Correlation Matrix[10]	0.0281	0.435	0.42
DFA[11]	0.0237	0.493	0.49
Statistical Jump Model[12]	0.0269	0.451	0.44
Ours	0.0218	0.528	0.54

Overall, the proposed method outperforms all baseline models across all evaluation metrics, demonstrating strong predictive capability and clear advantages in asset return modeling. Specifically, in terms of MSE, the method achieves the lowest error of 0.0218, indicating higher accuracy in numerical regression of asset

returns. In contrast, traditional methods such as the Text-Based Correlation Matrix and the Statistical Jump Model show larger errors. This suggests that these models suffer from information loss or temporal lag when modeling complex macro signals and market responses.

The R^2 metric further confirms the structural modeling ability of the method in explaining return variations. With a coefficient of determination of 0.528, the proposed model leads all other approaches. This indicates that the joint design of macro text factors and time windows effectively captures the underlying drivers of asset prices. Traditional models, while showing some fitting ability during certain periods, fail to model dynamic semantic shifts and transitions in market states. As a result, their explanatory power remains limited.

In terms of Information Ratio (IR), the proposed model again performs best, reaching a value of 0.54. This shows that the model not only achieves high prediction accuracy but also maintains return stability and strategy feasibility. This advantage stems from the time window mechanism that captures delayed market responses and from the semantic encoder that provides a deep understanding of macro policy expressions. These components enable the model to extract stable predictive signals under volatile conditions.

This paper further gives the impact of text encoding dimension setting on regression accuracy, and the experimental results are shown in Figure 2.

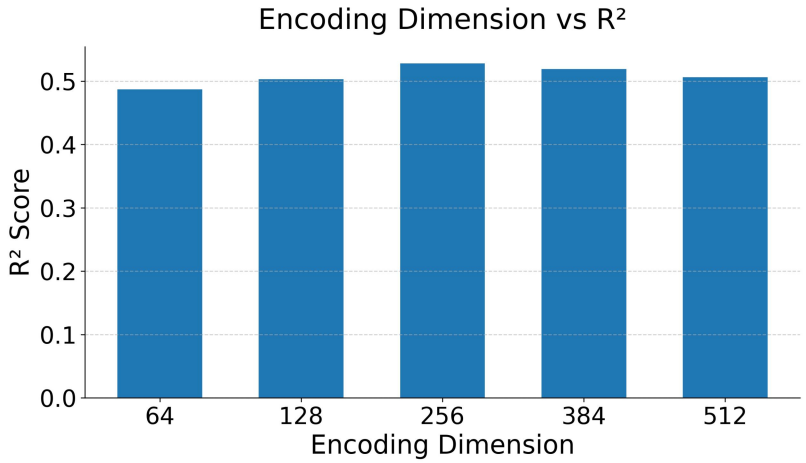

Figure 2. The impact of text encoding dimension setting on regression accuracy

The experimental results show that the change in text encoding dimension has a significant impact on the performance of the regression model. When the dimension is low (such as 64 or 128), the R^2 score is relatively low. This indicates that the model's semantic representation ability is limited and cannot fully capture the deep information and structural signals in macroeconomic texts, which weakens its explanatory power for asset returns. As the dimension increases, model performance improves. At 256 dimensions, the R^2 score reaches its peak and outperforms all other settings.

This trend reflects the relationship between the expressiveness of textual factors and the depth of modeling. Properly increasing the semantic vector dimension helps the model better represent complex content such as policy signals and economic expectations. This enhances the alignment between semantic information and market response. However, overly high dimensions (such as 384 or 512) may introduce redundant features in practice. This can increase the risk of overfitting or diluting feature weights, which leads to a slight performance decline.

The results confirm the sensitivity of the proposed semantic modeling module to dimensional choices. They also highlight the need to balance semantic representation and generalization ability when designing macro text-driven regression frameworks. Choosing an appropriate encoding dimension is not only essential for

prediction accuracy but also critical to the model's robustness and practical applicability in financial market settings.

From the perspective of asset allocation modeling, setting the semantic vector dimension properly is fundamental for high-quality return forecasting. A suitable dimension allows the model to more effectively extract key information from unstructured policy texts. When combined with a time window mechanism, it further enhances the temporal continuity of return modeling and improves control over strategic responses. This provides technical support for building structure-aware and dynamic asset allocation strategies.

This paper also gives an evaluation of cross-market adaptability under the condition of asset class transformation, and the experimental results are shown in Figure 3.

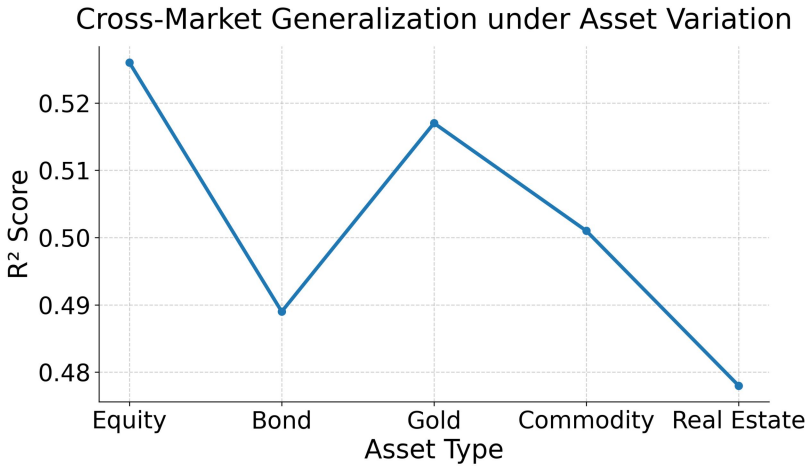

Figure 3. Evaluating cross-market adaptability under asset class changes

The experimental results show the regression performance of the proposed model across different asset classes, highlighting its adaptability in cross-market settings. As shown in the figure, the model achieves relatively high R^2 values on Equity and Gold assets, reaching 0.526 and 0.517 respectively. This suggests that when market signals align closely with macro texts or when assets are highly liquid, the model can more effectively capture the predictive value of textual factors.

In contrast, the model performs slightly worse on Bond and Real Estate assets. This is especially true for Real Estate, where the R^2 drops to 0.478. This indicates that the model faces limitations in capturing patterns for low-frequency or structurally illiquid assets. These markets may respond to macro texts with stronger delays, or the influence of text-driven signals may be weaker, which reduces the model's explanatory power.

For Commodity assets, the model shows moderate regression performance. It maintains a reasonable level of prediction accuracy in markets that are sensitive to global policies and exhibit high volatility. The model can extract semantic signals related to supply and demand or geopolitical risks from policy texts. However, its ability to handle unstructured information transmission still has room for improvement. This paper also presents a modeling stability analysis under the heterogeneity of input corpus sources. The experimental structure is shown in Figure 4.

The experimental results show that different input text sources have a clear impact on the regression performance of the model. This highlights both the disruptive and enhancing effects of textual heterogeneity on modeling stability. For single-source inputs such as "News Only" and "Policy Only," the R^2 values are 0.492 and 0.478 respectively. These results indicate that under limited information conditions, the model struggles to construct a complete semantic context and market response path. The regression becomes less stable and lacks sufficient explanatory power.

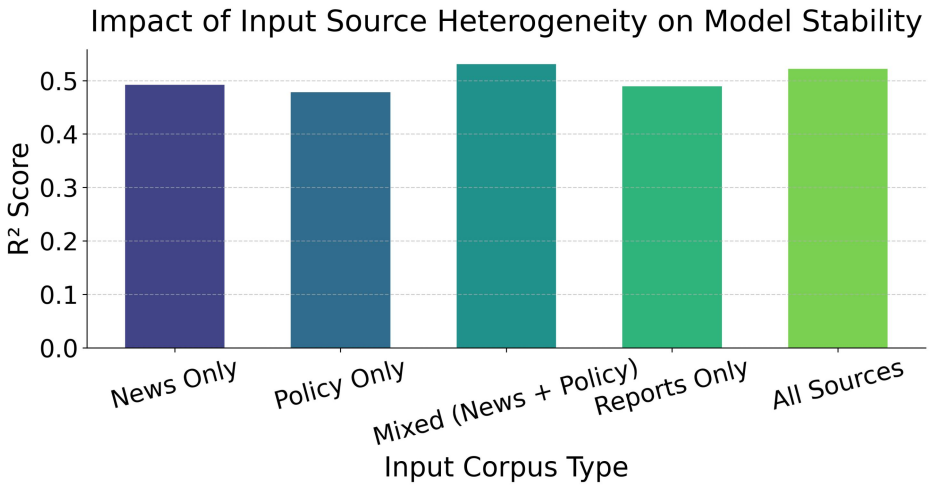

Figure 4. Analysis of modeling stability under the heterogeneity of input corpus sources

Notably, when the input uses the "Mixed (News + Policy)" combination, the R^2 rises to the highest value of 0.531. This shows that multi-source information fusion can enhance the representation of macro signals and improve the model's understanding of asset return drivers. It suggests that different text types provide complementary information. News captures short-term sentiment and event dynamics, while policy texts convey institutional direction and expectations. Combining the two helps form richer and more temporally structured factor representations.

The "Reports Only" group performs slightly better than policy texts but still falls short of the mixed-source configuration. This may be due to the delayed updates in research reports or the dominance of structural descriptions over dynamic information, making it harder for the model to track real-time signal fluctuations. In the "All Sources" condition, the model maintains a high R^2 value of 0.522. This further confirms the positive effect of textual diversity on model stability and predictive performance.

In summary, the experiment validates that the choice of input text type plays a critical role in building macro text-driven return regression models. Introducing heterogeneous sources and integrating them structurally not only improves the semantic completeness of the model but also strengthens its robustness and generalization in dynamic financial environments. This underscores the essential role of text source selection and integration strategies in financial modeling. This paper also gives the impact of changes in text noise intensity on the accuracy of return prediction, and the experimental results are shown in Figure 5.

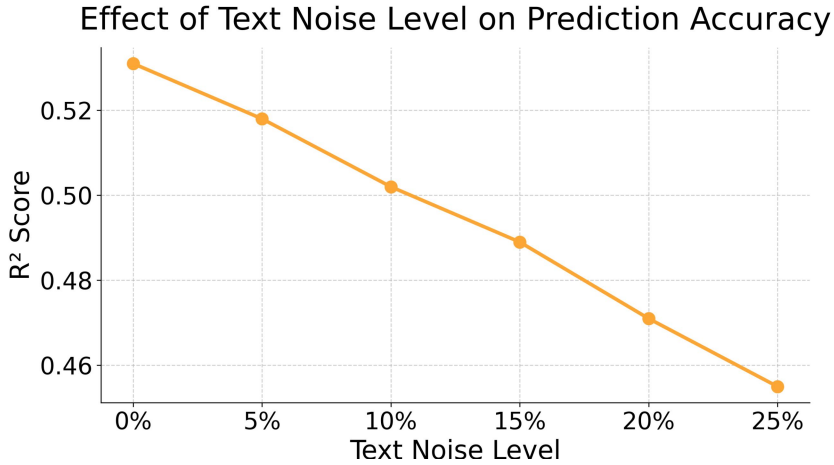

Figure 5. The impact of text noise intensity changes on return prediction accuracy

The experimental results clearly show the significant impact of text noise intensity on the regression performance of the model. As the noise ratio increases from 0% to 25%, the R^2 score drops steadily from 0.531 to 0.455. This indicates that stronger semantic disturbance in the text leads to weaker explanatory power of the model for asset returns. The trend highlights the importance of semantic integrity in constructing macro-driven factors. Text noise directly disrupts the coherence of semantic structures and the transmission of policy logic, making it difficult for the model to capture effective market signals.

The model remains relatively stable under 5% and 10% noise levels, with only slight decreases in R^2. This suggests that the proposed modeling framework has a certain degree of robustness under light perturbations. This robustness may come from the time window mechanism and residual feature fusion strategy, which help reduce the impact of individual text errors on overall regression performance. However, when the noise ratio exceeds 15%, model performance begins to decline significantly. This indicates that the noise has already damaged the core structure of the input features, making it harder for the model to maintain accurate predictions.

It is worth noting that under 25% noise intensity, model performance drops sharply. This shows that when too many invalid words or incorrect semantics are present in the text input, the model can no longer recover meaningful information from the context. This sets a higher requirement for macro text modeling. It must not only support structural extraction but also include noise reduction mechanisms or confidence assessment strategies to maintain prediction accuracy under uncertain conditions.

6. Conclusion

This paper proposes a regression algorithm for asset allocation returns that integrates macro text factors with temporal structure modeling. The goal is to address the limitations of traditional methods in handling unstructured information and their lack of dynamic response mechanisms. The proposed framework extracts deep semantic features through a text encoder and captures the market's phased response to macro signals using a time window mechanism. It enables continuous prediction of asset returns. The overall method is end-to-end and can handle challenges such as semantic redundancy, temporal lag, and multi-source heterogeneous input in macroeconomic environments. It demonstrates strong modeling stability and adaptability across asset types.

By introducing structured temporal modeling, this study enhances the model's ability to identify delayed effects of macro text influence. The predictive framework responds not only to short-term market fluctuations but also provides a logical explanation for medium- and long-term trends. This joint modeling of causal pathways and time structures offers more resilient and forward-looking support for dynamic asset allocation. Experimental results show that the proposed method outperforms several mainstream baselines in both regression accuracy and interpretability under various data conditions and noise disturbances. The method has strong generalizability and empirical value.

This research expands the application boundaries of financial text factors in asset allocation scenarios. It also provides a feasible paradigm for cross-modal fusion modeling in financial markets. In the context of information overload and rising financial uncertainty, the model's structural awareness, robustness, and adaptability have important practical implications. Especially under frequent macro policy changes and the rapid evolution of asset classes, the method offers key algorithmic support for building intelligent, structured, and dynamically optimized asset management systems.

7. Future work

Future research can explore integration methods for more diverse and heterogeneous text sources to improve the model's resilience under extreme market conditions. The framework may also be extended by incorporating risk factor modeling and multi-task learning, enabling applications in both return prediction and risk assessment. In addition, to improve the interpretability of the model structure, techniques such as graph

neural networks and causal graphs can be introduced. These enhancements can support more transparent and robust modeling of financial decision-making logic and promote the adoption of data-driven asset allocation in intelligent investment research and financial regulation.

References

[1] Baitinger E. Forecasting asset returns with network‐based metrics: A statistical and economic analysis[J]. Journal of Forecasting, 2021, 40(7): 1342-1375.

[2] Dahlquist M, Ibert M. Equity return expectations and portfolios: Evidence from large asset managers[J]. The Review of Financial Studies, 2024, 37(6): 1887-1928.

[3] Cohen G. Algorithmic trading and financial forecasting using advanced artificial intelligence methodologies[J]. Mathematics, 2022, 10(18): 3302.

[4] Dunbar K, Owusu-Amoako J. Cryptocurrency returns under empirical asset pricing[J]. International Review of Financial Analysis, 2022, 82: 102216.

[5] Dai Z, Li T, Yang M. Forecasting stock return volatility: the role of shrinkage approaches in a data‐rich environment[J]. Journal of Forecasting, 2022, 41(5): 980-996.

[6] Shen Z, Wan Q, Leatham D J. Bitcoin return volatility forecasting: A comparative study between GARCH and RNN[J]. Journal of Risk and Financial Management, 2021, 14(7): 337.

[7] Ge W, Lalbakhsh P, Isai L, et al. Neural network–based financial volatility forecasting: A systematic review[J]. ACM Computing Surveys (CSUR), 2022, 55(1): 1-30.

[8] Cong L W, Tang K, Wang J, et al. Deep sequence modeling: Development and applications in asset pricing[J]. arXiv preprint arXiv:2108.08999, 2021.

[9] Kisiel D, Gorse D. Portfolio transformer for attention-based asset allocation[C]//International Conference on Artificial Intelligence and Soft Computing. Cham: Springer International Publishing, 2022: 61-71.

[10] Nakayama Y, Sawaki T, Furuya I, et al. Text-Based Correlation Matrix in Multi-Asset Allocation[J]. arXiv preprint arXiv:2405.14247, 2024.

[11] Zhang Z, Wang Z, Ji Y, et al. Dynamic evolution of spatial distribution of energy factor allocation efficiency: Industrial sector in China[J]. Environment, Development and Sustainability, 2024: 1-19.

[12] Najjarpour A, Rostami M. Jump test and Estimate the Size and Probability of Jump in the Stock Market Using Stochastic Volatility Models[J]. Journal of Econometric Modelling, 2022, 7(1): 71-96.