Received: 16 April 2026 | Revised: 28 May 2026 | Accepted: 3 June 2026 | Published online: 5 June 2026

Journal of Finance and Financial Management, Volume 2, 2026
https://h-tsp.com/

Multimodal LLM Forecasting: Aligning News Semantics with Price Dynamics in Commodity Markets

Xiaoxiao Deng

DePaul University, Chicago, United States, College of Computing and Digital Media

Abstract: ***Theobjective of this study is to construct a large-language-model (LLM)-driven, text-enhanced time series forecasting framework in which unstructured news information is transformed into informative exogenous variables for futures price prediction. Unlike conventional pipelines that rely on bag-of-words statistics and shallow topic/sentiment mining, we leverage the contextual semantic understanding and reasoning capabilities of LLMs to extract thematic and sentiment signals from a large corpus of futures-related news headlines. Specifically, each headline is encoded into a high-dimensional semantic embedding by a pretrained LLM, from which fine-grained topic intensity and directional sentiment (bullish / bearish / neutral with strength) are derived and fed into the predictor as exogenous features. This paper addresses two critical design questions: why headlines over full articles, and why futures news over crude-oil-only news. First, news headlines act as highly condensed summaries that encapsulate the most decision-relevant information; for an LLM, headlines further mitigate context-length cost while preserving the core semantic and emotional cues, which is consistent with the headline-based topic-and-sentiment extraction adopted by Li et al. [1–5]. Second, we select the broader futures-news domain rather than crude-oil news alone because crude-oil-specific headlines are scarce, and because gold, natural gas and crude-oil futures exhibit well-established cross-commodity dependencies. To exploit this, we inject these empirical dependencies into the LLM as domain priors via a retrieval-augmented generation (RAG) module: the model dynamically retrieves established findings—e.g., Sujit & Kumar (2011), who show that gold-price fluctuations affect the WTI index and that countries' crude-oil dependence influences exchange rates and thus the purchasing power of gold, and Villar & Joutz (2006), who report that a 20% temporary shock to WTI exerts a 5% contemporaneous impact on natural-gas prices [6–9]—so that the extracted textual features implicitly carry cross-commodity transmission knowledge. This RAG-based prior injection both compensates for the scarcity of single-commodity news and enriches the information density of the resulting exogenous variables.***

Keywords: PSO-SVR hybrid model; Machine learning; Uncertainty sentiment; Empirical asset pricing.

1. Introduction

Considering the decay effect of indices, we construct a daily sentiment-strength index based on a large language model (LLM) rather than a conventional BERT classifier. The rapid development of social media has multiplied the channels through which people publish and consume messages, reflecting a wide spectrum of emotions and attitudes. Sentiment analysis, a key text-mining technology, employs computational linguistics to identify, extract, and quantify the affective information embedded in text. Whereas the conventional BERT framework [18–24] relies on a fixed multi-layer transformer fine-tuned on labelled data to output a sentiment-polarity probability in [0, 1], its limited parameter scale and pretraining corpus constrain its ability to fully comprehend the nuanced, domain-specific, and often implicit sentiment carried by short financial headlines. To overcome this, we adopt an instruction-tuned

LLM as the sentiment encoder: through carefully designed prompt engineering and a small set of in-domain instruction-following examples, the LLM performs zero-/few-shot sentiment scoring on each headline and emits a continuous polarity score between 0 and 1, where lower scores denote more bearish/negative sentiment and higher scores denote more bullish/positive sentiment. Crucially, the LLM is prompted to perform chain-of-thought reasoning before scoring, so that the polarity reflects not merely surface lexical cues but the headline's contextual market implication. The daily sentiment strength is then obtained by aggregating (e.g., averaging) the LLM-derived sentiment scores of all news headlines published on that day.

2. Preliminary and Algorithm Process

The textual feature-engineering module of our framework departs from the classical statistical topic-modeling lineage (LDA → NMF/SeaNMF) and is instead built upon a large language model (LLM)-driven, multimodal topic-and-sentiment representation pipeline. In this section we first recall the classical models as conceptual baselines, and then formalize our LLM-based replacement.

2.1 LDA to LLM-based Neural Topic Modeling

LatentLatent Dirichlet Allocation (LDA) is a groundbreaking generative probabilistic model that has long shaped natural language processing and machine learning. Topic modeling is a form of unsupervised learning that discovers abstract topics within large volumes of text, and is particularly useful for understanding the themes permeating large document collections such as digital libraries, news archives, and online forums [32–35]. However, LDA—and statistical topic models in general—represents each document as a bag of words and therefore ignores the semantic relationships and contextual ordering between words; for short, noisy financial headlines this assumption is especially fragile, and LDA's performance is further known to degrade as the number of topics grows.

To overcome these limitations, we replace LDA with an LLM-based neural topic-modeling paradigm. Rather than estimating word–topic and topic–document distributions from co-occurrence counts, each headline is encoded into a contextual semantic embedding by a pretrained large language model—optionally adapting a decoder-only LLM into a strong text encoder—so that semantically similar headlines lie close in the embedding space. These embeddings are then dimensionality-reduced and clustered to induce latent topics, after which a generative LLM produces human-readable, semantically coherent topic labels and assigns each headline a probabilistic membership over the topic set. This embedding-then-generate scheme inherits the topic-discovery objective of LDA while encoding contextual semantics, polysemy, and implicit meaning that the bag-of-words assumption discards, thereby transforming statistical topics into semantic topics.

2.2 From SeaNMF to a Multimodal LLM Topic–Sentiment Encoder

SeaMNF SeaNMF (Semantics-assisted Non-negative Matrix Factorization for short-text topic modeling) improves upon LDA by capturing word–context relations within a small window via skip-gram modeling, negative sampling, and a low-rank (nuclear-norm) factorization of the term–context matrix, factorizing it into lower-dimensional matrices that represent topics and their document-level weights, and thereby alleviating the sparsity and ambiguity of short texts [36–39]. While effective, SeaNMF remains a count-/co-occurrence-driven linear factorization: it relies on a manually constructed vocabulary, a fixed context window, and shallow statistical regularization, and consequently cannot model long-range dependencies, domain-specific jargon, irony, or implicit market sentiment in financial headlines.

We therefore generalize SeaNMF into a multimodal LLM topic–sentiment encoder that jointly produces the daily topic-strength and sentiment-strength signals consumed by the downstream predictor. The three classical techniques of SeaNMF are re-cast in the LLM paradigm as follows:

- From skip-gram windows to global self-attention. Instead of capturing word–context relations within a small, fixed window, the LLM's self-attention mechanism models long-range, bidirectional dependencies across the entire headline, yielding context-aware token and sentence representations that subsume and extend skip-gram semantics.

- From negative sampling to contrastive / instruction-tuned representation learning. Rather than contrasting true word occurrences against randomly sampled negatives, we leverage the LLM's pretrained semantic space—optionally refined by contrastive representation learning and lightweight instruction tuning—to separate relevant from irrelevant content, providing far stronger discrimination between salient and noisy headlines.

- From nuclear-norm low-rank factorization to embedding-space soft clustering. In place of explicitly minimizing the nuclear norm of a term–context matrix to obtain a low-rank topic structure, we obtain a naturally low-dimensional, semantically organized manifold directly from the LLM embeddings, on which soft clustering recovers the most prominent topics without hand-tuned L1/L2 regularization.

Finally, to exploit the fact that futures prices are driven jointly by textual news and numerical price series, the encoder is embedded in a cross-modality alignment framework: the LLM-derived semantic/sentiment embeddings (the text modality) and the historical price series (the numerical modality) are projected into a shared representation space and aligned via cross-modal attention, so that the most informative news signals are entangled with—and conditioned on—the price dynamics. This multimodal design upgrades SeaNMF's single-modality, statistics-only topic strength into a semantically rich, market-aware, and jointly-modeled exogenous feature, which is then passed to the forecasting module.

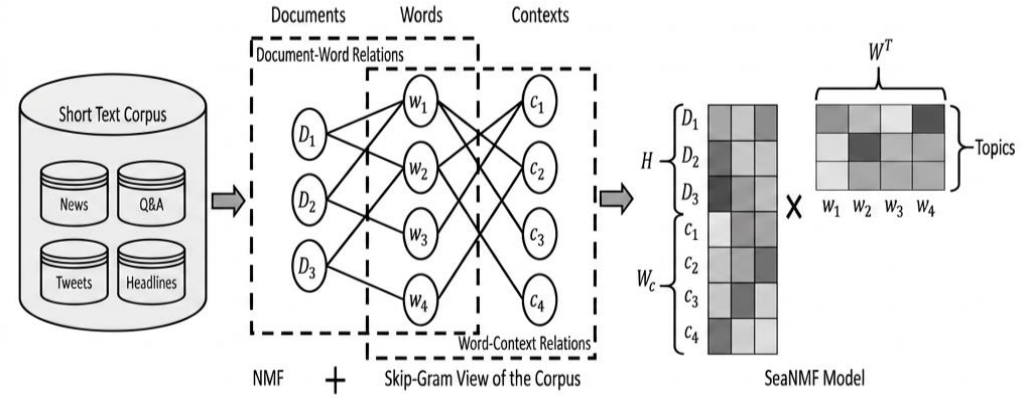


Figure 1

3. Simulation Experience

The objective of this experiment is to construct a semantic representation space from a dataset of news headlines, which is subsequently exploited for text-mining tasks—specifically, comparing the topic-modeling performance of an LLM-based neural topic model against the classical SeaNMF and LDA baselines. Unlike the conventional pipeline, which builds a discrete integer-indexed vocabulary and

feeds count-based representations into a statistical factorizer, our approach bypasses explicit vocabulary construction entirely: each headline is mapped directly into a high-dimensional contextual embedding by a pretrained large language model, and topics are induced by dimensionality reduction and soft clustering over these embeddings, followed by a generative-LLM topic-representation step.

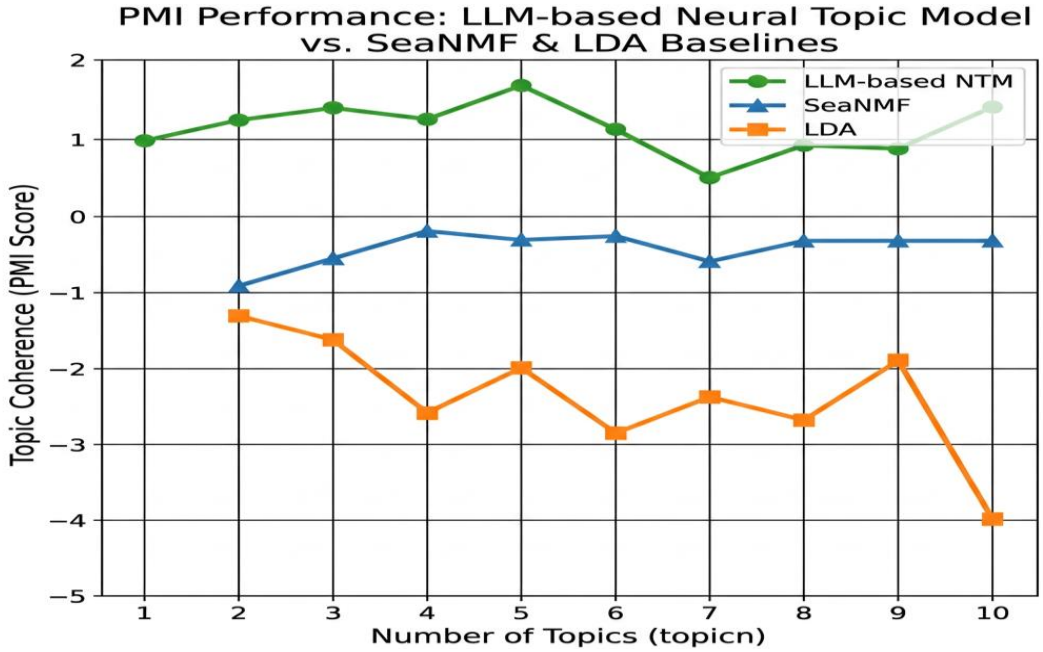


Figure 2

Figure 2. Across various numbers of topics, the LLM-based topic model attained consistently higher and more stable topic-coherence scores than both SeaNMF and LDA, indicating that LLM semantic embeddings capture the underlying thematic structures of the corpus more faithfully than count-based factorization or bag-of-words generative models. As depicted in Figure 2, the LLM curve dominates the SeaNMF and LDA curves and exhibits markedly lower variance as k varies.

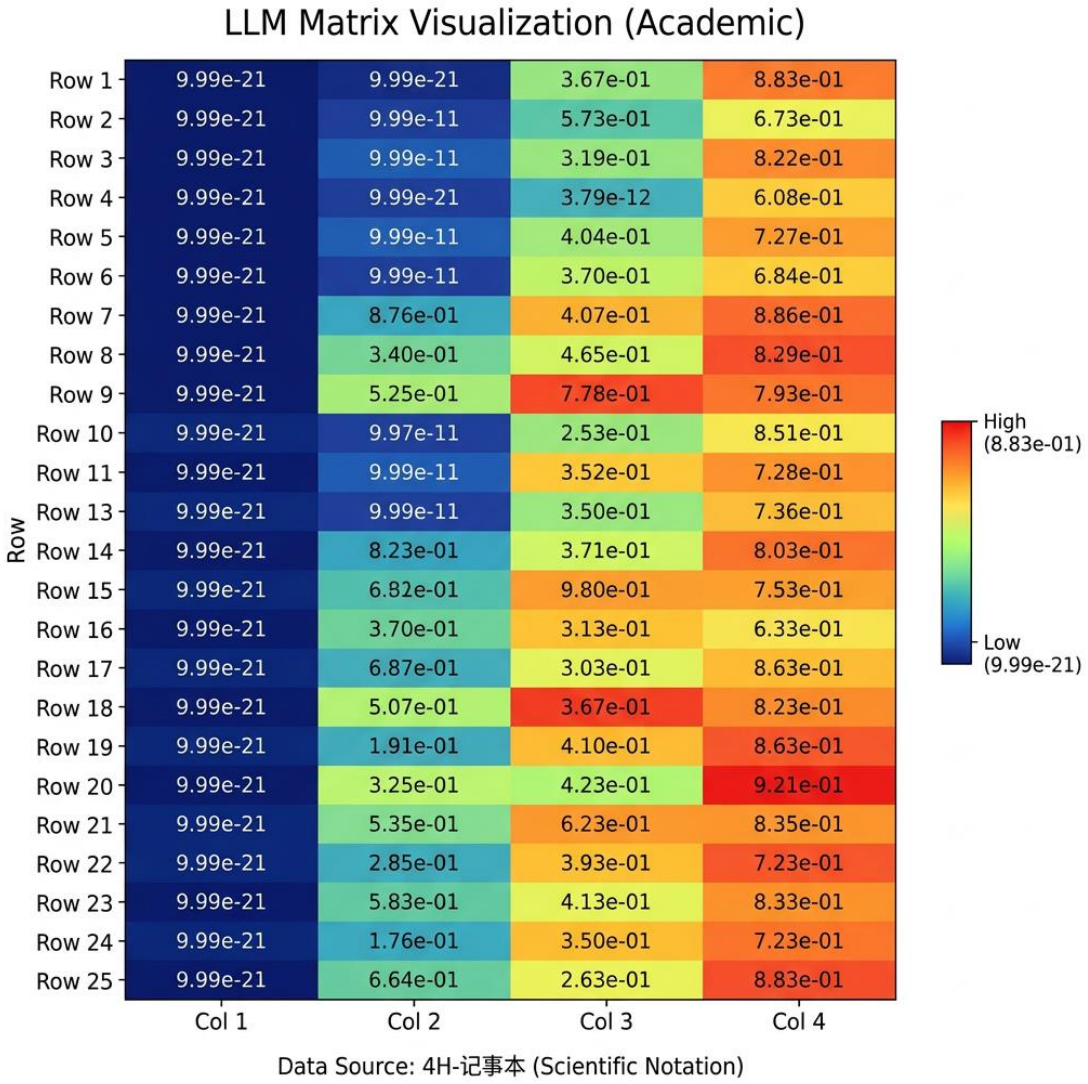
The optimal number of topics was determined to be four, corresponding to the highest coherence score; this choice is further validated by the distinct and semantically meaningful themes the LLM recovers, which align precisely with the commodities of interest: crude oil, gold, natural gas, and new-energy sources [40–43]. Whereas SeaNMF and LDA represent each topic merely as a ranked list of high-weight terms, the generative LLM additionally produces a concise, human-readable label and summary for each of the four topics, and the top-10 representative keywords it surfaces per topic confirm the model's ability to isolate clearly delineated themes within the textual data. The interconnections highlighted across these themes reflect the complex cross-commodity relationships within the market.

```
# ——— Proposed: LLM contextual embedding ———
import numpy as np
from sentence_transformers import SentenceTransformer # or any LLM encoder

encoder = SentenceTransformer("llm-embedding-model")
headlines = [line.strip() for line in open(investing_news, "r")]

# one forward pass → dense semantic embeddings (no vocab2id, no counting)
embeddings = encoder.encode(headlines, normalize_embeddings=True)
```

```
np.save(args.embedding_file, embeddings)
```



With the optimal number of topics established, the SeaMNF model was utilized to generate a 4H matrix representing the distribution of the four most relevant topics across the dataset. The conversion of the H matrix into a probability distribution, as shown in the code snippet, allows for a more nuanced understanding of the prevalence of each topic within the dataset.

$$\text{rmse} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

$$\text{mae} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|.$$

$$\text{mape} = \frac{100\%}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

4. Conclusions

The conclusions drawn from this study signify a substantial advancement in the application of large language model (LLM)-driven text-mining techniques for time-series forecasting, particularly within the commodities market. The integration of textual features—topic and sentiment signals derived from news headlines through an LLM—has proven to be a pivotal asset in enhancing the predictive accuracy of forecasting models [44–46]. This research has successfully constructed a time-series forecasting

framework that leverages the LLM-extracted thematic and sentiment information from news headlines to forecast future market trends, deeply integrating the relational semantics of text into the prediction pipeline.

The experimental results have conclusively demonstrated the superiority of the LLM-based topic model over both the classical SeaNMF and the traditional LDA in the context of short-text data such as news headlines. Whereas SeaNMF mitigates short-text sparsity through skip-gram modeling and negative sampling, and LDA relies on a bag-of-words generative assumption, the LLM encodes each headline into a contextual semantic embedding that natively captures long-range dependencies, polysemy, domain-specific jargon, and implicit market sentiment. Consequently, the LLM-based scheme achieves higher and more stable topic-coherence scores across various numbers of topics, yielding more accurate and robust topic extraction than either statistical baseline.

The determination of the optimal number of topics ($k = 4$) for the LLM-based model has been a critical finding of this study. This value not only corresponds to the highest coherence score but also aligns with the distinct and semantically meaningful themes relevant to the commodities market—namely crude oil, gold, natural gas, and new-energy sources. Beyond ranking representative keywords, the generative LLM produces human-readable labels and summaries for each of these four themes, underscoring the model's effectiveness in capturing—and explaining—the thematic structures embedded within the dataset.

While this study has made significant strides in applying LLM-based text mining to time-series forecasting, several avenues for future research remain. These include: (i) exploiting retrieval-augmented LLMs (RA-LLM) and instruction tuning to further sharpen domain-specific financial sentiment understanding; (ii) integrating real-time / streaming news for intra-day trading strategies, with LLM-based generative agents reasoning over breaking events as they unfold [47,48]; (iii) expanding the dataset to a broader range of commodities and financial instruments; and (iv) advancing toward a multimodal LLM forecasting architecture that jointly aligns textual news embeddings with numerical price series via cross-modal attention, thereby coupling qualitative market signals with quantitative dynamics.

In conclusion, this research has successfully demonstrated the potential of LLM-derived textual features in enhancing time-series forecasting models. The LLM's superior performance in extracting and interpreting meaningful topics from news headlines, coupled with the construction of daily LLM-based topic-strength and sentiment-strength indices, positions this study at the forefront of financial text analytics. The implications of these findings for empirical asset pricing are profound, offering a new, semantically grounded dimension for understanding and predicting market movements.

References

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp.529-551, April 1955.

[2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] Qian, C., Guo, Y., Mo, Y., & Li, W. (2025). WeatherDG: LLM-Assisted Procedural Weather Generation for Domain-Generalized Semantic Segmentation. IEEE Robotics and Automation Letters, 10(6), 5919–5926. https://doi.org/10.1109/lra.2025.3559821

[4] Qian, C., Guo, Y., Li, W., & Markkula, G. (2025). Weathergs: 3D Scene Reconstruction in Adverse Weather Conditions Via Gaussian Splatting. In 2025 IEEE International Conference on Robotics

and Automation (ICRA) (pp. 185–191). IEEE. 2025 IEEE International Conference on Robotics and Automation (ICRA). https://doi.org/10.1109/icra55743.2025.11128699

[5] Wu, W., Guo, Y., Li, Qi, & Jia, C. (2024). Exploring the potential of large language models in identifying metabolic dysfunction-associated steatotic liver disease: A comparative study of non-invasive tests and artificial intelligence-generated responses. Liver International, 45(4). https://doi.org/10.1111/liv.16112

[6] Qian, C., Li, W., Guo, Y., & Markkula, G. (2025). WeatherEdit: Controllable Weather Editing with 4D Gaussian Field (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2505.20471

[7] Guo, Y., Qian, C., Mo, Y., & Sangpetch, A. (2025). GaussianSlicer: Efficient Surface Reconstruction from Cross-sectional Slices with Gaussian Splatting. In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1–5). IEEE. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp49660.2025.10890834

[8] J. Li, T. B. Culver, P. P. Persaud, and J. M. Hathaway, "Developing nitrogen removal models for stormwater bioretention systems," Water Research, vol. 243, p. 120381, 2023.

[9] J. Li and T. B. Culver, "Review of process-based nitrogen model for agricultural fields with implications for nitrogen simulations in stormwater BMPs," Environmental Modelling & Software, vol. 151, p. 105363, 2022.

[10] Li, T. B. Culver, C. R. Burgis, W. Zhang, and J. A. Smith, "Validating Nitrogen Removal Models with Field Bioretention Data," Journal of Environmental Engineering, vol. 150, no. 8, p. 04024037, 2024.

[11] Li, "Nitrogen Removal Models for Stormwater Bioretention Systems," Ph.D. dissertation, University of Virginia, 2023.

[12] Liang, J., Wang, Z., Ma, Z., Li, J., Zhang, Z., Wu, X., & Wang, B. (2024). Online training of large language models: Learn while chatting. arXiv preprint arXiv:2403.04790.

[13] Wang, Z., Su, J., Zhou, M., Zeng, H., Jia, M., Lv, X., ... & Zhang, D. (2025). SheetBrain: A Neuro-Symbolic Agent for Accurate Reasoning over Complex and Large Spreadsheets. arXiv preprint arXiv:2510.19247.

[14] Ge, W., Wang, Z., Wang, P., Liang, J., Cai, Z. G., Mai, Z., & Wang, B. (2024). Towards Gamifying Interactive Language Learning using Large Language Models for Children.

[15] Wang, Z., Zhang, Q., Liu, T., & Li, C. (2024). Analyzing Financial News Sentiment with NLP to Forecast Market Trends. International Journal of Engineering and Management Research, 14(5), 6-11.

[16] Rao, Jiarui, et al. "Optimizing Stock Market Return Forecasts with Uncertainty Sentiment: Leveraging LLM-based Insights." Proceedings of the 2024 5th International Conference on Big Data Economy and Information Management. 2024.

[17] Rao, Jiarui, and Jionghao Lin. "Ramo: Retrieval-augmented generation for enhancing moocs recommendations." arXiv preprint arXiv:2407.04925 (2024).

[18] Rao, Jiarui, Qian Zhang, and Xinqiu Liu. "Applications Analyzing E-commerce Reviews with Large Language Models (LLMs): A Methodological Exploration and Application Insight." Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023 7.01 (2024): 207-212.

[19] Zhang, Qian, et al. "Sea MNF vs. LDA: Unveiling the power of short text mining in financial markets." International Journal of Engineering and Management Research 14.5 (2024): 76-82.

[20] Rao, Jiarui, et al. "Integrating Textual Analytics with Time Series Forecasting Models: Enhancing Predictive Accuracy in Global Energy and Commodity Markets." Innovations in Applied Engineering and Technology (2023): 1-7.

[21] Zhang, Qian, and Jiarui Rao. "Enhancing Financial Forecasting Models with Textual Analysis: A Comparative Study of Decomposition Techniques and Sentiment-Driven Predictions." Innovations in Applied Engineering and Technology (2022): 1-6.

[22] Lin, Jionghao, et al. "Automatic large language models creation of interactive learning lessons." European Conference on Technology Enhanced Learning. Cham: Springer Nature Switzerland, 2025.

[23] Peng, Jingyang, et al. "Automated bias assessment in ai-generated educational content using ceat framework." International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2025.

[24] Rao, Jiarui, and Qian Zhang. "Deconstructing Digital Discourse: A Deep Dive into Distinguishing LLM-Powered Chatbots from Human Language." Journal of Theory and Practice in Education and Innovation 2.2 (2025): 18-25.

[25] Deng, Xiaoxiao. "Enhancing Neural Network Performance on Tabular Data via Knowledge Distillation and RankGauss Transformation." 2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2025.

[26] Deng, Xiaoxiao. "Graph Inference Towards ICD Coding." 2025 3rd International Conference on Artificial Intelligence and Automation Control (AIAC). IEEE, 2025.

[27] Zi, Yun, and Xiaoxiao Deng. "Joint modeling of medical images and clinical text for early diabetes risk detection." Journal of Computer Technology and Software 4.7 (2025).

[28] Xu, Ting, et al. "Clinical NLP with attention-based deep learning for multi-disease prediction." 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC). IEEE, 2025.