HKUST BUSINESS SCHOOL 香港科大商學院 | WORLD CLASS IN ASIA

Advancing the Last Research Mile with Domain-Specific Modeling: Designing Novel Machine Learning Artifacts for Financial Data Analytics

Summary (498 words):

The prevalence of multi-modal data in the financial market—ranging from firm reports and news articles to earnings call recordings—presents both an opportunity and a challenge for decision-makers. Traditional analytical models often struggle to extract actionable insights from this vast, complex landscape, motivating the design of novel machine learning artifacts specifically tailored for multi-modal and unstructured financial data. Key design challenges include the multi-modality of financial data, where textual transcripts, accompanying audio from earnings calls, and firm fundamentals each contain unique and complementary information. Additionally, financial text is typically lengthy, low in signal-to-noise ratio, and filled with industry-specific jargon that carries context-specific meanings. Downstream financial analytics tasks also present unique characteristics, with financial textual similarity often hinging on subtle semantic differences, and financial sentiment analysis being highly susceptible to distribution shifts.

To address the challenges in financial analytics modeling, we have designed a suite of novel machine learning artifacts for multi-modal financial data integration and analysis (Figure 1):

1. **DeepVoice**: We developed DeepVoice [1], a predictive analysis system that leverages both textual data (quarterly earnings call transcripts) and audio data (call recordings). Grounded in Mehrabian's communication theory, which posits that vocal cues and their congruence with verbal information are essential in communication, DeepVoice is an end-to-end deep learning model designed specifically to address challenges of time variation, vocal-verbal integration, and vocal complexity within multi-modal financial data.

2. **FinBERT**: To adapt language modeling to the finance domain, we designed FinBERT [2], a large language model pretrained on a substantial corpus of finance-specific text. FinBERT integrates financial knowledge to deliver precise contextual insights and consistently outperforms traditional methods, including the Loughran and McDonald dictionary, in financial text analysis. To maximize the impact of this research, we have publicly released the FinBERT model, which has achieved more than 1 million monthly downloads on HuggingFace (Support document 5). FinBERT has also been piloted internally by leading financial institutions, including the Hong Kong Monetary Authority, Goldman Sachs Asia, and AllianceBernstein (Support document 6-8).

3. **Task-specific Financial Analytical Models**: We further developed specialized models that adapt natural language processing (NLP) techniques to specific financial analytics tasks, such as financial text similarity matching [3] and financial sentiment analysis under distribution shift [4]. These models account for unique financial data characteristics and distribution challenges in downstream analytics tasks.

Multi-modal model	NLP model	Task-specific model
DeepVoice (*Management Information Systems Quarterly 2023*)	FinBERT (*Contemporary Accounting Research 2023*)	Financial Text Similarity (*Finding of NAACL 2024*) Financial Sentiment Analysis under Distribution Shift (*EMNLP 2023*)

Table 1: Developing finance domain-specific machine learning artifacts.

Our work sheds light on design principles that can generalize to broader domains. First, we advocate for **"the last research mile" through domain adaptation**. While the fields of NLP and machine learning have been transformed by general-purpose AI models, our work demonstrates that these models still fall short in finance-specific tasks. Financial domain data exhibits unique characteristics, such as multi-modality, low signal-to-noise ratio and context relevant, which necessitate tailored artifacts that can bridge "the last research mile" through domain adaptation [5].

Second, we propose **innovative artifacts for multi-modal data integration**, emphasizing the need to harmonize diverse data types—such as text, audio, and numerical data—to extract complementary insights and improve predictive accuracy across complex analytical tasks.

Verification: This project is principally led by university-based faculty for R&D.

Supporting Documents:

- **Publications**:
    - (Support Document 1) Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. (2023). Unlocking the Power of Voice for Financial Risk Prediction: A Theory-Driven Deep Learning Design Approach. *Management Information Systems Quarterly*, *47*(1), 63-96.
    - (Support Document 2) Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806-841.
    - (Support Document 3) Liu, J., Yang, Y., & Tam, K. Y. (2024). Beyond Surface Similarity: Detecting Subtle Semantic Shifts in Financial Narratives. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2641-2652).
    - (Support Document 4) Guo, Y., Hu, C., & Yang, Y. (2023). Predict the Future from the Past? On the Temporal Data Distribution Shift in Financial Sentiment Classifications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1029-1038).
- **Industry Supporting Letters and Testimonials:**
    - (Support Document 5) Community impact (Github stars and Huggingface monthly downloads)
    - (Support Document 6) Supporting letter from Hong Kong Monetary Authority Chief Digitalisation Officer
    - (Support Document 7) Supporting letter from Head of Portfolio Management Group, Goldman Sachs Asia

References:

[1]. Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. (2023). Unlocking the Power of Voice for Financial Risk Prediction: A Theory-Driven Deep Learning Design Approach. *Management Information Systems Quarterly*, 47(1), 63-96.

[2] Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806-841.

[3] Liu, J., Yang, Y., & Tam, K. Y. (2024). Beyond Surface Similarity: Detecting Subtle Semantic Shifts in Financial Narratives. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2641-2652).

[4] Guo, Y., Hu, C., & Yang, Y. (2023). Predict the Future from the Past? On the Temporal Data Distribution Shift in Financial Sentiment Classifications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1029-1038).

[5] Nunamaker Jr, J. F., Briggs, R. O., Derrick, D. C., & Schwabe, G. (2015). The last research mile: Achieving both rigor and relevance in information systems research. *Journal of Management Information Systems*, *32*(3), 10–47.