Research project in Data Science

COM-412

Enhancing News-Based Asset Pricing with Information-Driven Trading Signals

Project Proposal

by Matthias Wyss (SCIPER 329884)
William Jallot (SCIPER 341540)

Master in Data Science
Minor in Financial Engineering
MA3

Prof. Pierre Collin Dufresne

Contents

Autumn Semester - 2025

1 Introduction

This document is a project proposal for the Data Science Master's research project which will take place during the 2025 autumn semester. It aims to investigate whether incorporating information-driven trading (ITI) can enhance the predictive power of news-augmented asset pricing models. This work builds on recent evidence that transformer-based text embeddings (e.g., FinBERT) improve return forecasts when incorporated into traditional factor models.

Recent studies have combined natural language processing techniques with asset pricing factor models to predict returns and have demonstrated significant predictive power [1]. However, these approaches omit a crucial element: information-driven trading (ITI), which can materially alter the value of public news. While decision tree methods have been used to detect ITI from various microstructural variables, only a few works have introduced a dedicated ITI metric, most notably the measure in [2]. That metric increases ahead of earnings releases, mergers, and other news announcements, and helps explain return reversals and broader asset pricing effects.

Our project aims to unite these two information channels by combining an embedding model (e.g., FinBert [3]) to create news embeddings with ITI scores, weighting each news signal according to the level of informed trading that precedes it. Concretely, we will explore ways to embed the ITI metric directly into the embedding representation so that the impact of a headline is automatically scaled by the prevailing level of informed trading.

2 Preliminary plan, work separation

The work will be divided so that each of us maintains a primary area of focus while actively supporting one another across tasks. Matthias will primarily handle the natural language processing component, which involves collecting financial news data, generating text embeddings using models such as FinBERT [3], and developing baseline pricing models that incorporate these embeddings. William will focus on the market microstructure dimension, working with microstructural data to compute ITI scores based on established methodologies, or potentially extending or refining the metric. These scores will be incorporated within the pricing framework and later integrated into the news embeddings. We will explore several modeling approaches and collaborate closely on model integration, experimental design, and the incorporation of traditional factor models.

3 Data Collection and Preliminary Sources

As preliminary resources, we have identified several datasets that are well suited to our analysis. This list is not exhaustive and will be expanded as the project progresses:

- Reuteurs financial news from 2006 to 2013 [4]
- Bloomberg and Reuters dataset [5]
- FNSPID [6]
- Nifty [7]
- FinSen [8]
- SEntFiN [9]

References

[1] Liao Zhu, Haoxuan Wu, and Martin T. Wells. *A News-based Machine Learning Model for Adaptive Asset Pricing*. 2021. arXiv: 2106.07103 [q-fin.ST]. URL: https://arxiv.org/abs/2106.07103.

[2] VINCENT BOGOUSSLAVSKY, VYACHESLAV FOS, and DMITRIY MURAVYEV. "Informed Trading Intensity". In: *The Journal of Finance* 79.2 (2024), pp. 903–948. DOI: https://doi.org/10.1111/jofi.13320. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13320. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13320.

[3] Dogu Araci. *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. 2019. arXiv: 1908.10063 [cs.CL]. URL: https://arxiv.org/abs/1908.10063.

[4] Dan Benayoun. *Processed Dataset of Financial News Articles from Reuters (2006-2013)*. Year of dataset release.

[5] Xiao Ding Philippe Remy. *Financial News Dataset from Bloomberg and Reuters*. https://github.com/philipperemy/financial-news-dataset. 2015.

[6] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. *FNSPID: A Comprehensive Financial News Dataset in Time Series*. 2024. arXiv: 2402.06698 [q-fin.ST].

[7] Raeid Saqur. "NIFTY-LM Financial News Headlines Dataset for LLMs". In: *ArXiv* (2024). URL: https://arxiv.org/abs/2024.5599314.

[8] Wenhao Liang, Zhengyang Li, and Weitong Chen. "Enhancing Financial Market Predictions: Causality-Driven Feature Selection". In: *arXiv e-prints* (2024), arXiv–2408.

[9] Ankur Sinha et al. "¡scp¿SEntFiN¡/scp¿ 1.0: ¡scp¿Entity-aware¡/scp¿ sentiment analysis for financial news". In: *Journal of the Association for Information Science and Technology* 73.9 (Mar. 2022), pp. 1314–1335. ISSN: 2330-1643. DOI: 10.1002/asi.24634. URL: http://dx.doi.org/10.1002/asi.24634.