Wharton
University of Pennsylvania

Jacobs Levy Equity
Management Center
for Quantitative Financial Research

Discussion: Expected Returns and Large Language Models

Sophia Zhengzi Li, Rutgers Business School

AI Application in Finance

Gu, Kelly, and Xiu (2020): Empirical Asset Pricing via Machine Learning

- 2050+ Google Scholar citation and counting
- Inspired a large body of follow-up work including my own:

  ML to predict volatility: ***Automated Volatility Forecasting*** *(MS forthcoming)*

  ML to predict correlation: ***Forecasting and Managing Correlation Risks*** *(Working Paper)*

This paper:

- Apply state-of-the-art LLMs to news for predicting returns
- Broad scope: 16 global equity markets and news articles in 13 languages

Summary

Models

- LLM (ChatGPT, LLAMA, LLAMA2, RoBERTa, BERT)
- Word-based Models (Word2vec, SESTM, LMMD)
- ML Models (RIDGE, LASSO, RF, NN)

Sentiment Analysis

- Textual features from news to predict binary outcome (1 if ret>0), compare model performance on 1) sentiment prediction accuracy, 2) return predictive power using news sentiment

Return Prediction

- Estimate RIDGE using one-day-ahead ret as dep var and textual features from a model as inputs
- Use predicted returns as sorting variables

Summary

Main Findings

- Returns respond slowly to news
- LLMs outperform traditional word-based models in sentiment analysis and return prediction

Overall

- Extremely comprehensive (big data, tons of analyses, efforts for writing three papers into one)
- Highly educational (excellent details on models and implementation), recommend to everyone
- Expect similar impact to that of Gu, Kelly, and Xiu (2020) in the years to come

Comment 1: Sentiment Score

Primary aim of sentiment analysis:

delineate relation between specific text-based features $x_{i,t}$ and binary sentiment label $y_{i,t}$ on training articles: $E(y_{i,t}|x_{i,t}) = \sigma(x'_{i,t}\beta)$; $\sigma(x)$ is a logistic link function

- To achieve this, require a sentiment label for each article in the training sample

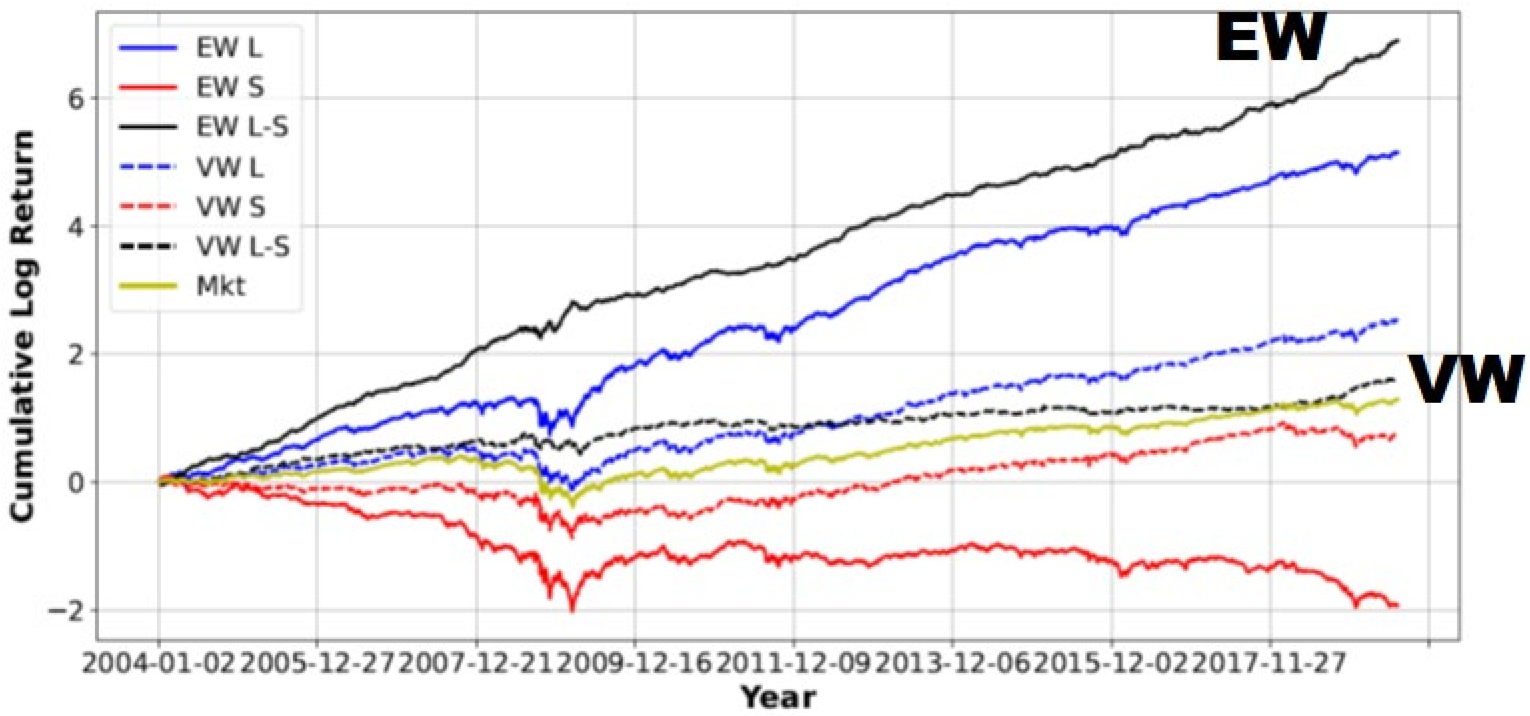
- Create sentiment labels based on 3-day returns surrounding the news article

Comment 1: Sentiment Score

3-day return as label, window too long

- 3-day return might not be sharp enough: confounding news or even non-news market-moving events
- Instead, high-frequency and instantaneous market reaction to news (i.e., 15-min) is more meaningful

Figure 4: One-day-ahead Portfolio Performance based on LLaMA2



Significantly lower L-S portfolio return for large firms, likely due to more confounding news and faster price reactions of large firms

Recommendation 1: use intraday high-frequency returns as labels

Comment 1: Sentiment Score

Recommendation 2: Compare with manual labeling by expert readers

- **Bloomberg News Sentiment:** fine-tuned using datasets that include expert-labeled sentiments
- **Refinitiv (formerly Thomson Reuters News Analytics):** models are trained using a combination of machine learning and expert-labeled data
- **StockTwits:** platform aggregates user-generated content and assigns sentiment scores based on positive or negative mentions of stocks
- **RavenPack:** uses supervised learning methods that involve expert-labeled data as part of the training

Comment 2: Benchmark to News Mom of Jiang, Li, and Wang (2021)

Jiang, Li, and Wang (2021 JFE):

Pervasive underreaction: Evidence from high-frequency data

- 26 overnight and 15-min returns per day; return in interval j on day t as r_t^j, j = 1,2,…, 26
- Combine intraday firm-level news and return data to create news-driven returns

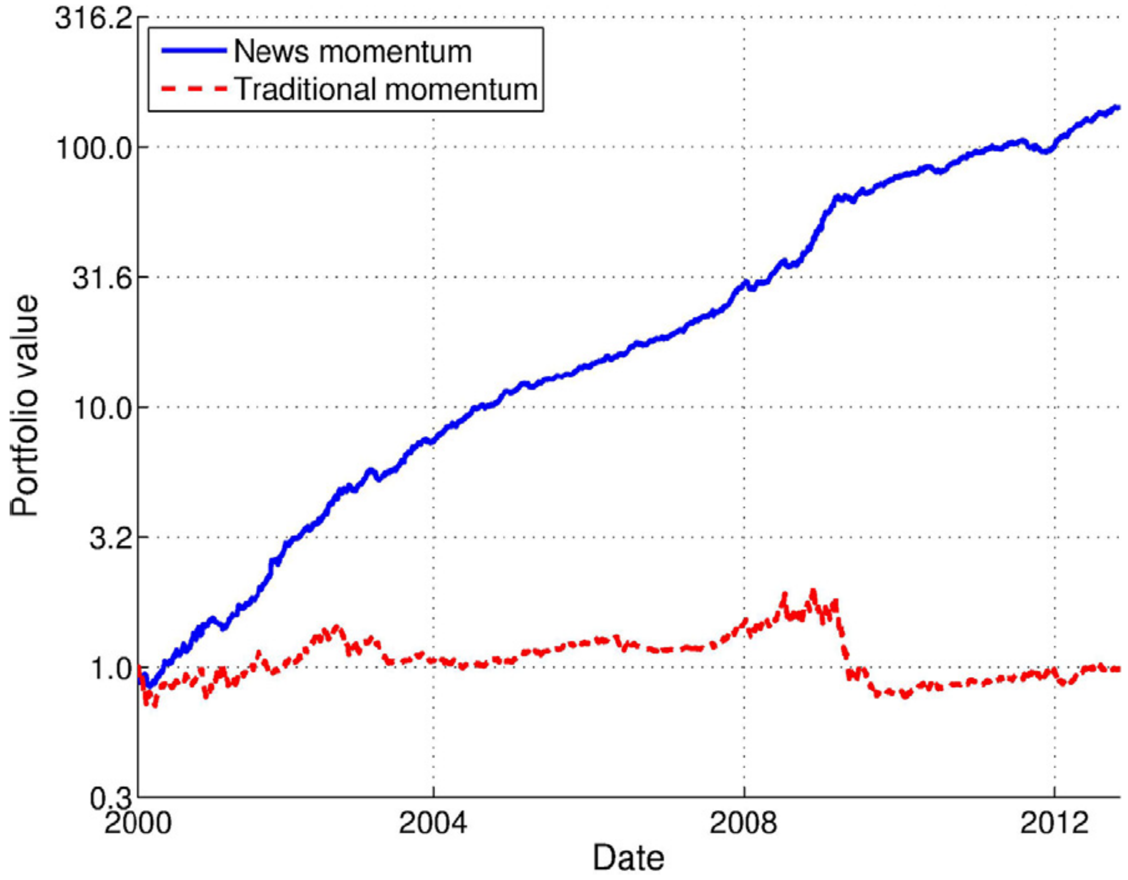
$$r_{t,news}^{j} = \begin{cases} r_t^j & \text{if there is a news story in interval } j, \\ 0 & \text{otherwise,} \end{cases}$$

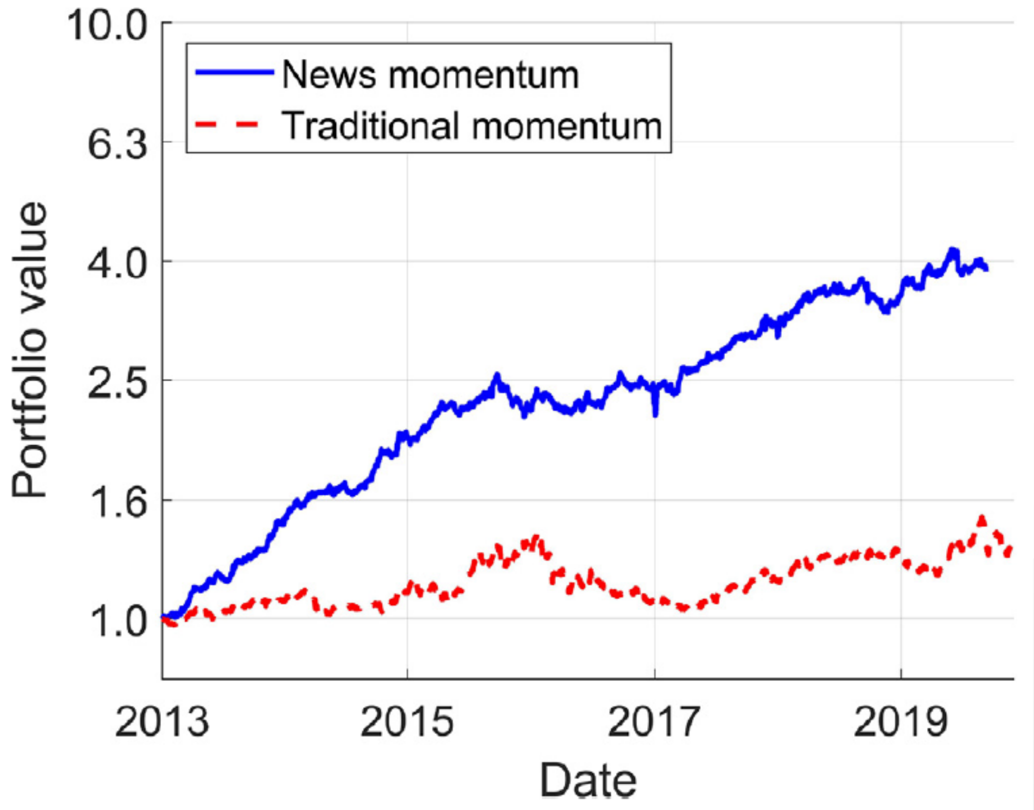
- Construct daily close-to-close "news return" as signals:

$$R_{t,news} = \prod_{j=1}^{26} (1 + r_{t,news}^{j}) - 1.$$

- One-day news-return signal to predict 5-day-ahead return

Comment 2: Benchmark to Jiang, Li, and Wang (2021)





- Over 2000-2019 sample ($1 price filter), EW annual return 34%, VW annual return 25%
- Survives transaction cost, a leading hedge fund still actively trades on it
- Easy to implement yet very effective, serves as a benchmark

Comment 3: Is Predictability Front-Loaded?

- Use news between 9am on day t-1 to 9am on day t to predict 9am-9am return on day t+1

- Assuming daily TC of 10 bps for large stocks and 20 bps for small stocks, net (after TC) annual Sharpe Ratio is around 1.5

- What if the predictability is front-loaded, i.e., concentrated at the beginning of the holding period, such as 1-min after 9am, where immediate trading is difficult? The predictability of large and liquid stocks might be more front-loaded. It might be worth conducting a more granular analysis at the intraday level.

Comment 4: News Category and News Clustering

Examine news separately across various categories

- Likely stronger predictive power based on fundamental news

Explore news clustering effects

- Test whether holding period return based on previous day's news signal is driven by the momentum of news (i.e., good news is followed by good news, and vice versa)

Conclusion

- Extremely comprehensive and highly educational, impactful in the years to come
- Very well written and very enjoyable to read, highly recommend it to everyone
- Further analyses: comparisons to alternative sentiment scores, benchmarking to news-mom strategy, exploring possible front-loaded return predictability, and examining news categories and news clustering
- I look forward to reading future versions!