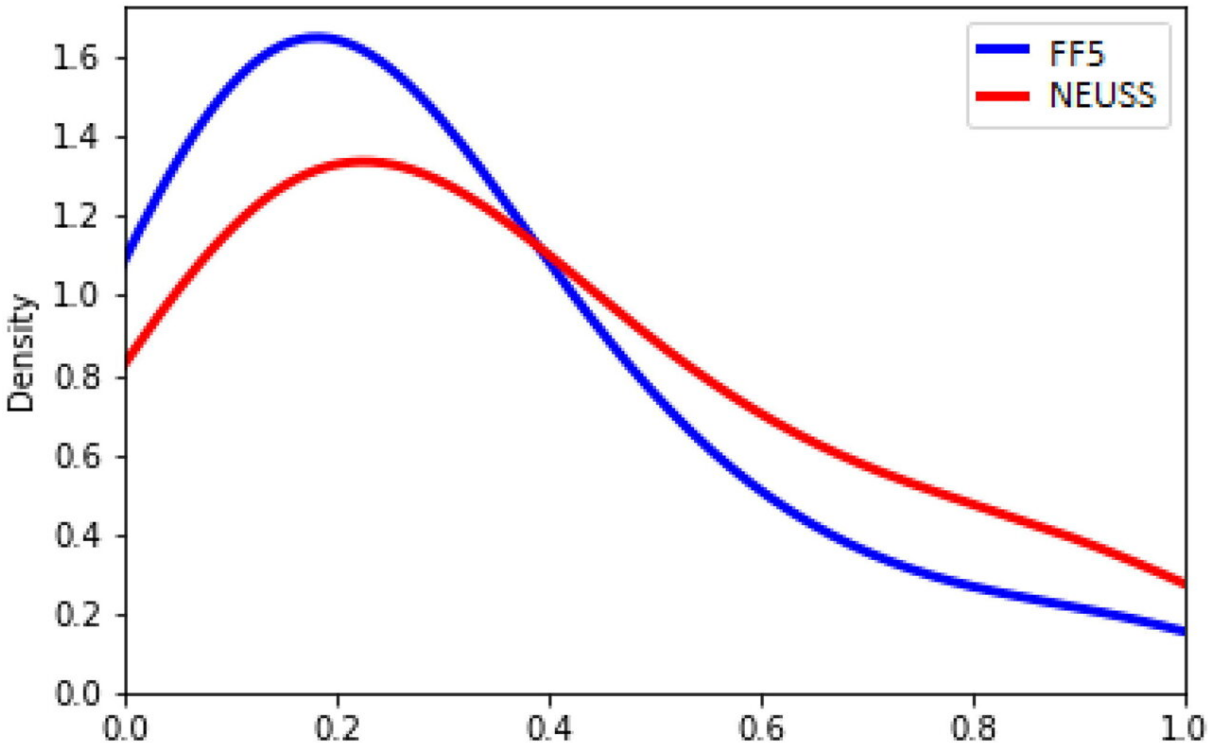
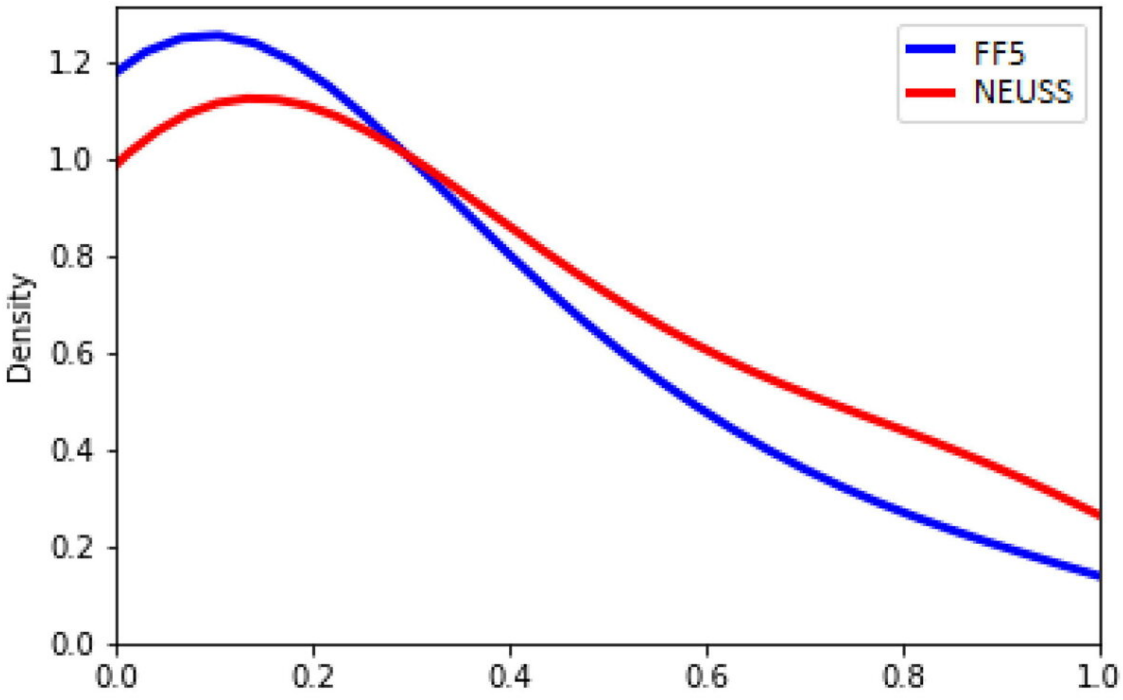
Data scientists predict stock returns with AI and online news

July 13 2023, by Louis DiPietro



(a) In-sample Adjusted R-squared



(b) Out-of-sample Adjusted R-squared

Adjusted R-Squared for explanation for FF5 vs NEUSS model. The plot illustrate the density adjusted R-squared for FF5 vs NEUSS for in-sample explanation and out-of-sample explanation comparisons. As seen, NEUSS outperforms FF5 in both in-sample and out-of-sample explanation. (a) In-sample Adjusted R-squared. (b) Out-of-sample Adjusted R-squared. Credit: *Data Science in Science* (2023). DOI: 10.1080/26941899.2023.2187895

For years, the financial press has helped inform investors of all stripes. Cornell researchers have discovered it can also inform the algorithm behind a new financial predicting model.

In their paper, "News-Based Sparse Machine Learning Models for Adaptive Asset Pricing," published in *Data Science in Science* in April, the researchers draw from interdisciplinary fields such as machine learning, natural language processing (NLP) and finance to build a new, interpretable machine-learning framework that captures stock- and industry-specific information and predicts financial returns with greater accuracy than traditional models.

"One of the knocks on machine learning is it's not interpretable," said Martin Wells, the Charles A. Alexander Professor of Statistical Sciences in the Cornell Ann. S Bowers College of Computing and Information Science and the paper's senior author. "Often when researchers use big models such as these, they may not know what the outputs mean or what is underlying the model. This research leverages text data from the news to build interpretable machine-learning models where you can see the important features explicitly."

The text helps with "clustering the data," bringing order to the chaotic results algorithms can produce, said lead author Liao Zhu, Ph.D. '20, who started working in the finance industry after finishing the paper. "Our hypothesis is that the financial news could do better in helping us better understand what type of stocks are related to certain tradable assets."

These assets could include exchange-traded funds (ETF), a bundle of stocks that tracks an entire sector, he said.

The paper is a continuation of Zhu's previous research that emerged from his early doctoral studies under Wells and Robert Jarrow, the Ronald P. & Susan E. Lynch Professor of Investment Management at the Samuel Curtis Johnson Graduate School of Management. Peter (Haoxuan) Wu, Ph.D. '23 is a co-author of the paper.

Applying traditional statistics methods to market data to explain stock returns is not new. Neither is using text data: Investors have used sentiment analysis, a subfield of natural language processing, to mine online text for positive or negative words associated with a company that, in theory, may signal a stock price's rise or fall.

The new research treads new ground by proposing a flexible prediction framework that bridges market data and text data without sentiment analysis, and integrates new, interpretable machine-learning algorithms. The researchers borrow the method of "word embeddings" from natural language processing and use an algorithm to create "asset embeddings" for a specific set of tradable assets from financial news. After converting both text and market data into numbers, the researchers then deploy custom-designed algorithms to crunch the numbers.

"Our algorithm is not using the sentiment from the news but using the news as guidance for what assets or words to consider for each specific

stock or industry, which reveals more stock- and industry-specific information," Zhu said.

To develop their models, researchers scraped a massive corpus of online financial news articles from 2013 to 2019 and fed it to their algorithm, which began mapping particular assets and words associated with specific stocks and industries. With an AI-optimized language map in hand, researchers had more insight into specific assets and words to consider.

Using this method, the team developed two separate models. The News Embedding UMAP Sparse Selection (NEUSS) model predicts returns for individual stocks, and the News Sparse Encoder with Rationale (INSER) model identifies important words for each specific industry before using them to predict industry returns more accurately.

For example, the NEUSS model may conclude from the financial news that an exchange-traded fund that tracks the semiconductor manufacturing sector is useful to predict the stock returns of a specific tech company, but may not be useful to predict returns of other stocks in, say, retail or wholesale. The INSER model may pick up the word "plant" as important for the energy industry, but this word may not be relevant for other industries like social media.

The hybrid, interpretable strategy worked. The NEUSS model beat out the traditional predictive benchmark—called the Fama-French 5-factor model—by 50%, while the INSER model beat the benchmark (without industry-specific information) by 10%.

The use of advanced machine-learning algorithms with different types of data is helping to revolutionize the finance field, Zhu and Wells said.

"I think the AI revolution in finance is already there," Zhu said, "and this

paper is moving an aspect of that revolution forward."

More information: Liao Zhu et al, News-Based Sparse Machine Learning Models for Adaptive Asset Pricing, *Data Science in Science* (2023). DOI: 10.1080/26941899.2023.2187895