Master’s Thesis Presentation

Xiaoran Cheng
Department of Statistics
The University of Chicago

“Large Language Models and Textual Analysis for Stock Return Prediction”

May 1, 2024, at 2:00 PM
Searle 236, 5735 S Ellis Avenue

Abstract

We investigate the potential of large language models (LLMs) to extract trading signals from textual data for predicting stock returns. Traditional financial textual analyses use numerical data or simple text-to-numeric embedding methods, such as the Bag of Words (BoW) method, which oversimplifies the complexity of human language and fails to capture semantic and syntactic meanings. To produce more accurate numerical representations of financial texts, we use more advanced language models called LLMs. LLMs are highly parameterized and are pre-trained on extensive datasets across various topics from human history. Specifically, this study leverages the Open Pre-trained Transformers (OPT) and Large Language Model Meta AI 2-13b (Llama2-13b) to embed financial data into numerical representations for econometric modeling. We evaluate the predictive ability of these embeddings for stock returns by using both ridge regression and neural networks. The dataset includes a comprehensive collection of news articles and alerts from Refinitiv's Thomson Reuters Real-time News Feed and Third Party Archive, matched to stock price data from CRSP and Datastream-EIKON, covering the period from January 1996 to June 2019. Our results demonstrate that LLM embeddings have high predictive power for stock market dynamics.