D I S S E R T A T I O N D E F E N S E

Lee Gao

Friday, April 15, 2016
3:00 pm
384 Posner Hall

Applications of Machine Learning and Computational Linguistics in Financial Economics

In the world of the financial economics, we have abundant text data. Articles on Wall Street Journals and Bloomberg Terminals, corporate SEC filings, earnings call transcripts, social media messages etc. contain rich information about the financial market and investor behaviors. Extracting meaningful signals from unstructured and high dimensional text data is not an easy task. However, with the development of machine learning and computational linguistic techniques, processing and statistically analyzing textual documents tasks can be accomplished within reasonable computation time nowadays, and many applications of statistical text analysis in social sciences have proven to be successful.

In my thesis, I conduct statistical text analyses using datasets constructed from the SEC corporate filings to retrieve information about the financial market macroeconomic conditions and investor behaviors. First, using the text data from the management discussions and analysis in corporate annual reports (10-K files), I examine whether the management discussions contain information that reveals a firm's exposure to systematic risk, and construct a risk factor based on textual information that can explain the cross-sectional variations in expected stock returns (Chapter 1). Second, using a unique text dataset containing letters to shareholders written by institutional investment managers, I analyze whether fund manager discussions contain insights in predicting returns of the market portfolio (Chapter 2). In addition to conducting empirical tests in asset pricing using textual data, I also construct a theoretical model to explain the interaction between corporate takeover activities and cash holdings behaviors, and I calibrate my model to the U.S. market mergers and acquisitions data (Chapter 3).

I demonstrate a variety of machine learning, natural language processing and dynamic programming techniques as powerful tools in complement to traditional econometric methods commonly adopted by economists. My work illustrates the potential of using text data as a new avenue for empirical research in financial economics. In particular, although computational linguistic techniques have made great achievements in social sciences such as political science and sociology, and they have also drawn a great deal of attention from financial industry practitioners, their applications in financial economics academia research is still limited. My work aims to fill this gap.