D I S S E R T A T I O N P R O P O S A L

Lee Gao

“Text Data: A New Avenue of Empirical Research in Financial Economics”

Friday, May 1, 2015
10:30 am
341-H Posner Hall

Well structured numerical data have long been the dominating source for empirical studies in financial economics. Although text financial data like journal articles, corporate regulatory filings, earnings call scripts and social media messages are more abundant than numerical ones, in terms of both amount and public availability, they have been utilized in quite limited number of financial economics researches. One of the main reasons is that text data are usually unstructured, fragmented and of very high-dimensional in nature, making the traditional data analysis tools familiar to financial economics researchers like regressions powerless. With the development of natural language processing and machine learning techniques in computational linguistics, analyzing text information in a systematical and efficient way becomes easier than ever and has attracted great attention from researchers in different disciplines, which also opens up a new avenue for empirical research in financial economics.

My research is dedicated to investigate and develop methodologies to efficiently extract financial market and corporate information contained in public available text data, and to understand their implications on market performance and corporate behaviors.

In the first chapter, I analyze the informativeness of text data contained in the management discussion and analysis section of SEC 10-K files about stock returns. I used the popular bag-of-words model in the computational linguistic literature to represent documents. In particular, each document can be seen as a list of tokens (the smallest text information unit, which usually can be words, phrases or several consecutive words depending on modeling) affiliated with their weights (can be counts, freqency, tf-idf etc.), assuming that the token position information is irrelevant. Although this assumption seems strong, it usually works very well. To summarize the stock return information contained in the high-dimensional text data into a one-dimensional text factor, I use multinomial inverse regression to project token count vectors onto a one-dimensional subspace that is most relevant to the stock returns. Then I run a cross-sectional linear regression to study the effects of text factor on stock returns, controlling existing well know covariates related to stock returns, including firm size, market-to-book ratio, stock turnover, NASDAQ dummy and industry dummy. I find that the text factor does have significant effects on stock returns. In the future, I plan to test the out-of-sample predictability of the text factor, and also examine the working mechanism of the text factor by studying its relationship with fundamental variables.

In the second chapter, I examine the factors that affect US stock market performance through a unique approach by checking topics in the discussions of investment management firm managers on their fund performance. The data are letters to shareholders extracted from SEC N-CSR(S) files, which are reports to shareholders mandatory for all the registered investment management firms in the US. I first

measure the sentiment of fund managers in a dictionary approach, meaning counting frequency of tokens of particular types (positive, negative, uncertain etc.) using pre-built word lists in literature. I find that negative words frequency is informative about market performance while positive and uncertain words frequency is not. Then I generate topics in the letters using a famous model in computational linguistics -- Latent Dirichlet Allocation, and find that frequency of different topics is also informative about market performance. In the future, I plan to examine the economic factor related to each text-based topic, so we can intuitively understand the driving force of market performance by looking at the topics. I will also check the out-of-sample predictabilities of sentiment and topic information.

In the third chapter, I investigate the relationship between corporate takeover activity and firm's cash holding policies. I develop a discrete-time infinite-horizon model with heterogeneous agents, solve the model numerically and calibrate it to US data on firm's takeover decisions and cash holdings. I find that market average cash holdings are increasing in acquisition opportunities as both acquirers and targets hold more cash than stand-alones. Acquirers hold more cash because of their larger motivations to avoid external financial cost. Targets hold more cash to attract acquirers as their cash holdings can be used by acquirers to reduce external financial cost in both current and future acquisitions, and the effect of reducing future acquisition cost originates from the model setup that firms have the option to make repeated acquisitions, which is not possible in a static model.