Textual Analysis, Volatility Forecasting, and the Cross-Section of Equity Option Returns

By Jie Cao, Bing Han, and Xintong Zhan

Keywords: Textual analysis; volatility forecasting; option return; machine learning; sentiments

Abstract

The proposed study explores the impact of new text-based measures of perceived volatility on the cross-section of equity option returns. The ever-increasing contents in digital text encode rich yet complex cues about the firms and provide an ideal source for understanding market participants' perceptions about stock volatility, the key input for option pricing. We use textual analysis and machine learning approach to extract new information regarding future stock volatility from various resources and examine its impact on equity option pricing.

Forecasting volatility is difficult and the error can lead to inaccuracy in option pricing (Figlewski (1989); Figlewski and Green (1999)). While statistical models like GARCH captures a good deal of the volatility dynamics, it ignores the rich information content available to the market participants (such as firm disclosures, management discussions, reports from analysts, opinions from investors forum) that could help shape their volatility forecasts in real time.

However, it is unclear whether such qualitative information affects investors trading behavior and asset prices. Particularly, to what extent it is incorporated in the equity options market and whether the answer depends on the sources. Managers' and financial analysts' perceived uncertainty is more likely to be associated with potential fundamental change in volatility. Such information, once incorporated into option price, leads to a change in the premium of option price relative to the underlying stock price. The opinions from the social networks, however, could be different. On one hand, literature documents crowd wisdom from the non-experts (Chen et al. (2014)). On the other hand, sentiment can add noise to volatility information.

This study applies textual analysis to obtain measures of perceived volatility from various resources including financial reports, analyst forecasts, press releases, news media, and investors social networks. We examine the impacts of the perceived volatility by different participants on the cross-section of delta-hedged equity option returns to understand the nature of the information. We also use machine learning techniques to better categorize the volatility information from various unstructured data sources.

The contribution of the study is two-fold. First, we extend the literature on textual analysis in finance, accounting and economic literature. Previous studies largely focus on directional information about stock price movement. We focus on the volatility information, which is equally important given the difficulty of forecasting volatility. Second, the study sheds light on the option pricing practices and on the cross-sectional option return predictability. Our study demonstrates the sources of information used by option market participants.

A. Big Data

A1. Sources for textual analysis ***(unstructured data)***

We consider various digital resources to extract the opinions from different market participants. The most direct way that managers communicate with investors is financial report. Corporate

filings like 10-K and 10-Q files will be the main sources to obtain perceived uncertainty from the managerial team. In addition, managers answer questions in conference calls. Unlike written reports, the speech and tones in conference call give more cues about managers' perception. So, we will also extract information from conference call transcripts, which can be accessed through Factiva. Furthermore, we will look at analyst reports, which can be accessed through Thomson Reuters, and collect financial media news from *Wall Street Journal* and *Dow Jones News*, etc. Moreover, we will use web crawler to collect data from internet sources, such as Twitter and investor forums (e.g., SeekingAlpha.com).

A2. Stock and option data ***(very large data sets)***

We obtain data from the stock and option markets. Following Cao and Han (2013) and Cao et al. (2017), data on single name equity options are from the OptionMetrics Ivy DB database. The database starts from January 1996 and contains information on the entire U.S. equity option market, including daily closing bid and ask quotes; open interest; volume; implied volatility; and Greeks such as delta, gamma, and vega, which are computed by Option-Metrics based on standard market conventions. We obtain stock returns, prices, and trading volume from the Center for Research on Security Prices (CRSP). The annual accounting data are obtained from Compustat. The quarterly institutional holding data are from Thomson Reuters (13F). The monthly analyst coverage and forecasts data are from I/B/E/S. The intra-day stock quotes and trades data are from Nasdaq TAQ database. The intra-day option trades data are from OPRA database. The option open-close and order type data can be purchased from CBOE Market Data Express (MDX). We will also include OPRA data in our analysis, which contains intra-day option trades and quotes data over 2005-2015 period. It will allows us to exploit the real trading of individual equity option at high frequency. This data is very large, about 2 TB before unzip. It requires high-performance computing resources.

B. Machine Learning

We plan to apply machine learning approach to extract volatility related information and potential pattern. Machines can be trained to read the text and identify uncertainty information, at the same time control human subjectivity, which is a common issue in textual analysis using dictionary or word list. For example, the support vector machine (SVM) approach will find a hyper-plane in the input space and classify the information with certain constraints. For example, Manela and Moreira (2017) adopt the same approach to measure news implied volatility (NVIX) and find NVIX can explain the time-varying risk premium of market. Other than SVM approach, the neural network is also very popular in finance studies, as it can handle large datasets, detects all the interactions of the works, and have high accuracy. For example, *Word2Vec*, a natural language processing with deep learning algorithm developed by Google, can convert words to vectors and classify the vectors into different spaces. We plan to use Python as computing language.

References:

Cao, Jie, and Bing Han, 2013, Cross-section of option returns and idiosyncratic stock volatility, *Journal of Financial Economics* 108, 231-249.

Cao, Jie, Bing Han, Qing Tong, and Xintong Zhan, 2017, Option return predictability, Working Paper, Revise & Resubmit at *Review of Financial Studies.*

Chen, Hailiang, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang, 2014, Wisdom of crowds: The value of stock opinions transmitted through social media, 2014, *Review of Financial Studies* 27, 1367-1403.

Figlewski, Stephen, 1989, Options arbitrage in imperfect markets, *Journal of Finance* 44, 1289-1311.

Figlewski, Stephen, and T. Clifton Green, 1999, Market risk and model risk for a financial institution writing options, *Journal of Finance* 54, 1465-1499.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137-162.