UCONN LIBRARY

University of Connecticut
OpenCommons@UConn

Doctoral Dissertations | University of Connecticut Graduate School

12-17-2015

Using Twitter to Analyze Stock Market and Assist Stock and Options Trading

Yuexin Mao
University of Connecticut, USA, yuexin.mao@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
Mao, Yuexin, "Using Twitter to Analyze Stock Market and Assist Stock and Options Trading" (2015). *Doctoral Dissertations*. 1000.
https://opencommons.uconn.edu/dissertations/1000

Using Twitter to Analyze Stock Market and Assist Stock and Options Trading

Yuexin Mao, Ph.D.

University of Connecticut, 2015

ABSTRACT

Twitter has rapidly gained popularity since its creation in March 2006. Stock is a popular topic in Twitter. Many traders, investors, financial analysts and news agencies post tweets about various stocks on a daily basis. These tweets reflect their collective wisdom, and may provide important insights on the stock market. In this dissertation work, we investigate using the tweets concerning Standard & Poor 500 (S&P 500) stocks to analyze the stock markets and assist stock trading.

The first part of the dissertation focuses on understanding the correlation between Twitter data and stock trading volume, and predicting stock trading volume using Twitter data. We first investigate whether the daily number of tweets that mention S&P 500 stocks is correlated with the stock trading volume, and find correlation at three different levels, from the stock market to industry sector and individual company stocks. We then develop two models, one based on linear regression and the other based on multinomial logistic regression, to predict individual stock trading volume into three categories: low, normal and high. We find that the multinomial logistic regression model outperforms the linear regression model, and it is indeed beneficial to add Twitter data into the prediction models. For the 78 individual stocks that have

significant number of daily tweets, the multinomial logistic regression model achieves 57.3% precision for predicting low trading volume and 67.2% precision for predicting high volume.

The number of tweets concerning a stock varies over days, and sometimes exhibits a significant spike. In the second part of the dissertation, we investigate Twitter volume spikes related to S&P 500 stocks, and whether they are useful for stock trading. Through correlation analysis, we provide insight on when Twitter volume spikes occur and possible causes of these spikes. We further explore whether these spikes are surprises to market participants by comparing the implied volatility of a stock before and after a Twitter volume spike. Moreover, we develop a Bayesian classifier that uses Twitter volume spikes to assist stock trading, and show that it can provide substantial profit. We further develop an enhanced strategy that combines the Bayesian classifier and a stock bottom picking method, and demonstrate that it can achieve significant gain in a short amount of time. Simulation over a half year's stock market data indicates that it achieves on average 8.6% gain in 27 trading days and 15.0% gain in 55 trading days. Statistical tests show that the gain is statistically significant, and the enhanced strategy significantly outperforms the strategy that only uses the Bayesian classifier as well as a bottom picking method that uses trading volume spikes.

In the third part of the dissertation, we investigate the relationship between Twitter volume spikes and stock options pricing. We start with the underlying assumption of the Black-Scholes model, the most widely used model for stock options pricing, and investigate when this assumption holds for stocks that have Twitter volume spikes. We find that the assumption is less likely to hold in the time period before a Twitter volume spike, and is more likely to hold afterwards. In addition, the volatility of a

stock is significantly lower after a Twitter volume spike than that before the spike. We also find that implied volatility increases sharply before a Twitter volume spike and decreases quickly afterwards. In addition, put options tend to be priced higher than call options. Last, we find that right after a Twitter volume spike, options may still be overpriced. Based on the above findings, we propose a put spread selling strategy for stock options trading. Realistic simulation of a portfolio using one year stock market data demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking account of commissions and ask-bid spread, while S&P 500 only increases 12.8% in the same period.

Using Twitter to Analyze Stock Market and Assist Stock and Options Trading

Yuexin Mao

M.S., University of Bridgeport, 2009

B.S., University of Electronic Science and Technology of China, China, 2007

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

Copyright by

Yuexin Mao

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Using Twitter to Analyze Stock Market and Assist Stock and Options Trading

Presented by

Yuexin Mao, B.S., M.S.

Major Advisor _____
Dr. Bing Wang

Associate Advisor _____
Dr. Mukul Bansal

Associate Advisor _____
Dr. Swapna Gokhale

Associate Advisor _____
Dr. Song Han

Associate Advisor _____
Dr. Mohammad Khan

University of Connecticut

2015

ACKNOWLEDGMENTS

I am deeply indebted to my major advisor, Dr. Bing Wang for her excellent guidance, encouragement and support through the entire duration of my Ph.D. study at UConn. Her abundant knowledge, sharp insights, extraordinary vision and outstanding passion have always guide me in the right direction. I express my heartfelt gratitude to her. I also would like to thank Dr. Wei Wei for his guidance. His excellence in research and perfection in academia have set an example of academic perfection, and have always been my strongest motivation to complete my Ph.D. study.

I am very fortunate to have Dr. Mukul Bansal, Dr. Swapna Gokhale, Dr. Song Han, and Dr. Mohammad Khan serve on my committee. It was my great pleasure working with such intelligent and responsible professors. I would like to thank Dr. Shengli Zhou, Dr. Jie Huang for their help on my work in network coding project.

I would like to extend my gratitude to my colleagues, Dr. Xian Chen, Dr. Yuan Song, Dr. Wei Zeng, Ruofan Jin, Yanyuan Qin and Levon Nazaryan, to name a few, for their great help to both my research and graduate life.

Last but not least, I would like to thank my family. Without them, none of the achievements I have made is possible. I am truly grateful to Shu, who has always been cheering me up and standing by me through the good times and bad. I want to give my deepest gratitude to my parents for their love, understanding, support and encouragement, and for letting me pursue my dreams far from home. To my family,

I dedicate this dissertation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction and Motivation

Twitter is a widely used online social media that enables users to send and read short 140-character messages called tweets. Users of Twitter can follow other users that they are interested in, post tweets that can be viewed by the public, retweet other users' tweets and even send them messages directly. Twitter has rapidly gained popularity since its creation in March 2006. As of September 2015, it has more than 500 million users, with more than 320 million being active users [66]. Twitter provides a light-weight, easy form of communication for users to share information about their activities, and for media to spread news. Topics in Twitter range from daily life to current events, breaking news, and others. The fast growth of Twitter has drawn much attention from researchers in different disciplines. Researchers have studied various aspects of Twitter. Existing studies on Twitter have investigated the general characteristics of the Twitter social network (e.g., [34], [42]) and the

social interactions within Twitter [32]. Several studies use tweets to predict real-world events such as earthquakes [56], seasonal influenza [2], the popularity of a news article [8], and popular messages in Twitter [31].

Stock market prediction has attracted much attention from researchers in both academia and business. In financial economics, the efficient-market hypothesis (EMH) (e.g., [26], [27]) states that stock market prices are largely driven by new information and follow a random walk hypothesis. The random walk hypothesis asserts that current market price fully reflects all available informations, implied that past and current information is immediately incorporated into stock prices, thus the price changes are only driven to new information or news. Since news is by definition unpredictable and random, thus, resulting price changes are unpredictable and random. However, several studies show that stock market prices do not follow a random walk (e.g., [28], [24], [16]) and can be predicted in some cases thereby challenging the assumptions of random walk hypothesis. Furthermore, although news may be unpredictable, early indicators can be extracted from Twitter to predict changes in stock market indicators [14, 44].

Stock is a popular topic in Twitter. Many traders, investors, financial analysts and news agencies post tweets about various stocks on a daily basis. These tweets reflect their collective wisdom, and may provide important insights on the stock market. Several studies have investigated predicting stock market using Twitter. Bar-Haim et al. [9] predict stock price movement by analyzing tweets to find expert investors and collect experts' opinions. Several studies use Twitter sentiment data to predict the stock market. Bollen et al. [14] find that specific public mood states in Twitter are significantly correlated with the Dow Jones Industrial Average (DJIA), and thus can be used to forecast the direction of DJIA changes. Zhang et al. [74] find that

emotional tweet percentage is correlated with DJIA, NASDAQ and S&P 500. Later on, Mao et al. [45] find that Twitter sentiment indicator and the number of tweets that mention financial terms in the previous one to two days can be used to predict the daily market return. Makrehchi et al. [44] propose an approach that uses event based sentiment tweets to predict the stock market movement, and develop a stock trading strategy that outperforms the baseline. In this dissertation, instead of considering the sentiment tweets on Twitter, we investigate the relationship between the number of tweets about stocks and stock market changes. Specifically, we investigate the correlation between the number of tweets about stock and stock trading volume, and further predict the stock trading volume using Twitter data. The number of tweets about stock sometimes exhibits a significant spike due to some events, which indicates a sudden increase of interests in the stock market. Motivated by the observation of Twitter volume spikes, we investigate when Twitter volume spikes occur and possible causes of Twitter volume spikes. Furthermore, we investigate whether Twitter volume spikes can be used to assist the stock trading and stock options trading.

1.2 Importance of Stock Trading Volume

The price of stocks are usually the primary interest for investors. After seeing the price of a stock, investors may next look into the data such as rate of return, market capitalization, earnings day or even ex-dividend date before considering the stock trading volume. Despite being ignored by many investors, stock trading volume is an important stock indicator and indeed has a relationship to stock price [40, 17].

Stock trading volume is the number of shares that are traded over a given period

of time, usually a day. Stock trading volume is treated as one of the most important stock indicators, and has a strong relationship with stock price [29]. First, stock trading volume indicates market liquidity, and the supply and demand for stocks. High trading volume of a particular stock indicates that this stock is more active in the stock market, and investors are placing their confidence in the investment. In contrast, low volume of a stock, even if it is rising in price, can indicate a lack of confidence among investors. Second, trading volume reflects pricing momentum. When stock trading volume is low, investors anticipate slower moving prices. When market activity goes up, pricing typically moves in the same direction. Last, trading volume can be treated as a sign of trend reversal. For example, a stock jumps 5% in one trading day after being in a long downtrend. To determine whether it is a sign of trend reversal for this stock, we can consider the trading volume. If the trading volume on the current day is high compared to the average daily trading volume several days before, it is a strong sign that the reversal is probably true. On the other hand, if the volume is relatively low, there may not be enough evidence to support a true trend reversal.

The random walk hypothesis asserts that past stock prices and trading volume can not be used to predict the future price changes and hence we can not rely on technical analysis to predict the future price returns. However, researchers believe that information contained in past stock prices is not fully incorporated in current stock prices, and hence, they believe that by observing the past stock prices, information can be obtained on future stock prices [36, 43]. Researchers believe that trading volume plays an important role to move the stock prices. Several studies have been made on trading volume and its relationship with stock returns (e.g., [40], [29], [17], [71]), suggesting that the price movements may be predicted by trading volume.

As the volume of tweets posted on Twitter about stocks increases, researchers are trying to find how the activity in Twitter data is correlated with time series from the stock market, specifically stock trading volume. The study [54] reports there indeed exists positive correlation between trading volume and the daily number of tweets for individual stocks. In this dissertation, we propose an in-depth analysis of the correlation between the number of tweets about stocks and stock trading volume at three different levels, from the stock market level to the industry sector level and then individual company stock level. Furthermore, we apply machine learning models to predict stock trading volume using Twitter data.

1.3 Twitter Volume Spikes and Stock Market

On April 23th, 2013, the Associated Press posted a tweet: "Breaking: Two Explosions in the White House and Barack Obama is injured.", which spreads quickly on Twitter platform, and exhibits a significant volume spike on this topic in a short amount of time. Although it has been confirmed that AP's official Twitter account has been hacked and the posted tweet was false soon, as shows in Fig. 1.3.1, S&P 500 index fell about 1% before quickly rebounding, briefly wiping out $136 billion US dollars followed by the false tweet. From this example, we notice that Twitter volume spikes have strong impact on stock market. Specifically, a tweet posted by an influential Twitter account is spreading quickly and can easily cause positive or negative reaction on stock market. On the other hand, when stock market has breaking news or important events, Twitter also reacts quickly and exhibits a significant volume spike. StockTwits [62] reports the average daily number of tweets mentioned

FIGURE 1.3.1: S&P 500 index fell 1% and quick rebounded in response to the high volume false tweets [13].

Apple's stock between October 27th and November 2nd has a spike around 14,000 in response to the event of Apple's quarterly earnings report released on October 27th. During the same period, Apple's stock price rose about 5%.

As stated before, most existing studies on the relation between Twitter and stock market focused on using Twitter sentiment data to predict the stock market return (e.g., [14], [74], [45], [44] [52]). In this dissertation, by analyzing Twitter volume spikes, we focus on whether they can shed light on the behavior of stock market, and whether the insights thus obtained can help to assist stock and stock options trading.

1.4 Contributions of This Dissertation

The contributions of this dissertation are three-fold: (i) analyzing the correlation between daily number of tweets and stock trading volume, and proposing modeling approaches to predict the stock trading volume, (ii) analyzing Twitter volume spikes related to S&P 500 stocks, and developing models to assist stock trading using Twitter volume spikes, and (iii) analyzing Twitter volume spikes related to S&P 500 stocks to find the relationship of Twitter volume spikes and stock options pricing.

First, we investigate the correlation between Twitter data and stock trading volume, and predict stock trading volume using Twitter data. More specifically, we investigate whether the daily number of tweets that mention S&P 500 stocks is correlated with the stock trading volume, and find correlation at three different levels, from the stock market to industry sector and individual company stocks. Our findings show that, the daily number of tweets related to S&P 500 stocks is correlated with stock trading volume at all three levels. We then develop two models, one based on linear regression and the other based on multinomial logistic regression, to predict individual stock trading volume into three categories: low, normal and high. We find that the multinomial logistic regression model outperforms the linear regression model, and it is indeed beneficial to add Twitter data into the prediction models. For the 78 individual stocks that have significant number of daily tweets, the multinomial logistic regression model achieves 57.3% precision for predicting low trading volume and 67.2% precision for predicting high volume.

Second, we investigate Twitter volume spikes related to S&P 500 stocks, and whether they are useful for stock trading. Through correlation analysis, we provide insight on when Twitter volume spikes occur. We find that Twitter volume spikes

often happen around earnings dates. Specifically, 46.4% of Twitter volume spikes fall into this category. We further explore whether these spikes are surprises to market participants by comparing the implied volatility of a stock before and after a Twitter volume spike. Our findings show that many Twitter volume spikes might be related to pre-scheduled events, and hence are expected to market participants. Furthermore, we investigate five possible causes of Twitter volume spikes including stock breakout points, large stock price fluctuation within a day and between two consecutive days, earnings days and high implied volatility. Our results show that only the last two factors show significant correlation with Twitter volume spikes. Moreover, we develop a Bayesian classifier that uses Twitter volume spikes to assist stock trading, and show that it can provide substantial profit. We further develop an enhanced strategy that combines the Bayesian classifier and a stock bottom picking method, and demonstrate that it can achieve significant gain in a short amount of time. Simulation over half a year stock market data indicates that it achieves on average 8.6% gain in 27 trading days and 15.0% gain in 55 trading days. Statistical tests show that the gain is statistically significant, and the enhanced strategy significantly outperforms the strategy that only uses the Bayesian classifier as well as a bottom picking method that only uses trading volume spikes.

Last, we investigate the relationship between Twitter volume spikes and stock options pricing. We start with the underlying assumption of the Black-Scholes model [12], the most widely used model for stock options pricing, and investigate when this assumption holds for stocks that have Twitter volume spikes. We find that the assumption is less likely to hold in the time period before a Twitter volume spike, and is more likely to hold afterwards. In addition, the volatility of a stock is significantly lower after a Twitter volume spike than that before the spike. We also find that implied

volatility increases sharply before a Twitter volume spike and decreases quickly afterwards. In addition, put options tend to be priced higher than call options. Last, we find that right after a Twitter volume spike, options may still be overpriced. Based on the above findings, we propose a put spread selling strategy for stock options trading. Realistic simulation of a portfolio using one year stock market data demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking account of commissions and ask-bid spread, while S&P 500 only increases 12.8% in the same period.

1.5 Dissertation Roadmap

The remainder of this dissertation is organized as follows. In Chapter 2, we describe our work on Twitter data and stock trading volume correlation analysis, and predicting stock trading volume using Twitter data. We first present the motivation of analyzing Twitter Data and S&P 500 Stocks in Section 2.1. We then describe data collection methodology and the datasets in Section 2.2. After that, we present correlation between Twitter data and stock trading volume in Section 2.3. Section 2.4 describes stock trading volume prediction using Twitter data. Finally, we summarize our work in Section 2.5.

In Chapter 3, we present the investigation of Twitter volume spikes related to S&P 500 stocks, and use Twitter volume spikes to assist stock trading. We first present the motivation of using Twitter volume spikes to assist stock trading in Section 3.1. We then describe data sets and define Twitter volume spike in Section 3.2. Section 3.3 presents the analysis of Twitter volume spikes and possible causes of these spikes.

Section 3.4 presents the trading strategies and their performance. Last, Section 3.5 summarizes our work in this chapter.

In Chapter 4, we investigate the relationship between Twitter volume spikes and stock options pricing. We first discuss the background and motivation in Section 4.1. Section 4.2 describes how we identify Twitter volume spikes. Section 4.3 briefly describes the lognormal stock price model and the Black-Scholes model. Section 4.4 analyzes the relationship between Twitter volume spikes and stock price model. Section 4.5 analyzes the relationship between Twitter volume spikes and stock options pricing. Section 4.6 presents a stock options trading strategy and evaluates its performance. Section 4.7 briefly discusses the choice of threshold for identifying Twitter volume spikes. Last, Section 4.8 summarizes our work.

Finally, we conclude this dissertation and present future work in Chapter 6.

Chapter 2

Correlating S&P 500 Stocks Trading Volume with Twitter Data

2.1 Introduction

Twitter is a widely used online social media. The fast growth of Twitter has drawn much attention from researchers in different disciplines. Researchers have studied various aspects of Twitter. Stock is a popular topic in Twitter, due to the real-time nature of tweets, researchers have become interested in using Twitter to predict stock market. Several studies research on the relation between Twitter and stock market focused on using Twitter sentiment data to predict the stock market return (e.g., [14], [74], [45], [44]). In this chapter, instead of focusing on sentiment, we investigate the correlation between the daily number of tweets that mention Standard & Poor 500 (S&P 500) stocks and S&P 500 stock trading volume. Our investigation is at three different levels, from the stock market, to industry sector, and then to indi-

vidual company stocks. We then develop two models, one based on linear regression and the other based on multinomial logistic regression, to predict individual stock trading volume into three categories: low, normal and high. It is useful to predict trading volume because when trading volume is high, it indicates that traders are interested in getting in or out the stock, so the stock can be easily traded and has high liquidity. On the other hand, trading volume being low indicates that the stock has a large bid-ask spread and is hard to trade. Our main findings are:

- We find that at the stock market level, the daily number of tweets that mention S&P 500 stocks is correlated with S&P 500 trading volume with correlation coefficient of 0.3. At the industry sector level, for six out of the ten GICS (Global Industry Classification Standard) industry sectors, there exists significant correlation between the number of daily tweets and the daily trading volume for the sector. In particular, Financials sector show the strongest correlations with correlation coefficient of 0.48. Last, at the individual company stock level, we investigate 78 individual stocks that have significant number of daily tweets, we observe that the number of daily tweets has strong correlation with stock trading volume at company stock level with median correlation coefficient of 0.53.

- We further develop two models, one based on linear regression and the other based on multinomial logistic regression, to predict individual stock trading volume into three categories: low, normal and high. We find that the multinomial logistic regression model outperforms the linear regression model, and it is indeed beneficial to add Twitter data into the prediction models. For the 78 individual stocks that have significant number of daily tweets, the multinomial

logistic regression model achieves 57.3% precision for predicting low trading volume and 67.2% precision for predicting high volume.

The rest of the chapter is organized as follows. Section 2.2 describes data collection methodology and the data sets. Section 2.3 presents the correlation between Twitter data and stock trading volume. Section 2.4 describes stock trading volume prediction using Twitter data. Last, Section 2.5 concludes the paper and presents future work.

2.2 Data Collection

2.2.1 Stock market data

We obtained daily stock market data from Yahoo! Finance [70] for the 500 companies in the S&P 500 list from February 16 , 2012 to May 31, 2013. At the stock market level, we consider the S&P 500 daily trading volume, which is the sum of the daily trading volume of 500 stocks in S&P 500 list.

At the sector level, we record the daily trading volume for each of the ten GICS sectors. GICS is an industry taxonomy developed by MSCI and S&P for use by the global financial community [68]. The GICS structure consists of ten industry sectors, including Information Technology, Financials, Consumer Discretionary, Consumer Staples, Industrials, Energy, Health Care, Materials, Telecommunications Services and Utilities. S&P 500 classifies each of the 500 companies into one of the ten industry sectors. For each sector, the daily trading volume of the sector is the sum of daily trading volume of all the companies in this sector.

At the company stock level, we focus on stocks that are more tweeted in S&P

TABLE 2.2.1: Number of companies and average number of tweets for the ten GICS sectors.

GICS sector	# of Companies	Avg. # of daily tweets
Information Technology	70	4451
Financials	81	1716
Consumer Discretionary	82	1649
Consumer Staples	41	783
Industrials	62	754
Energy	41	660
Health Care	51	540
Materials	29	406
Telecomm Services	8	292
Utilities	35	179

500. Specifically, we consider the individual stocks that has daily average number of tweets more than 25. Same as other two levels, we consider the daily trading volume for each individual stock.

2.2.2 Twitter data

In Twitter community, people usually mention a company's stock using the stock symbol prefixed by a dollar sign, for example, $AAPL for the stock of Apple Inc. and $GOOG for the stock of Google Inc. We use Twitter streaming API [67] to search for public tweets that mention any of the S&P 500 stocks using the aforementioned convention (i.e., putting a $ before the stock symbol). The reason why we use this convention is that some stock symbols are common words (e.g., A, CAT, GAS are stock symbols), and hence using search keywords without the dollar sign will result in a large amount of spurious tweets.

Fig. 2.2.1 plots the CCDF (complementary cumulative distribution function) of the average number of tweets for the S&P 500 stocks. We observe that the average

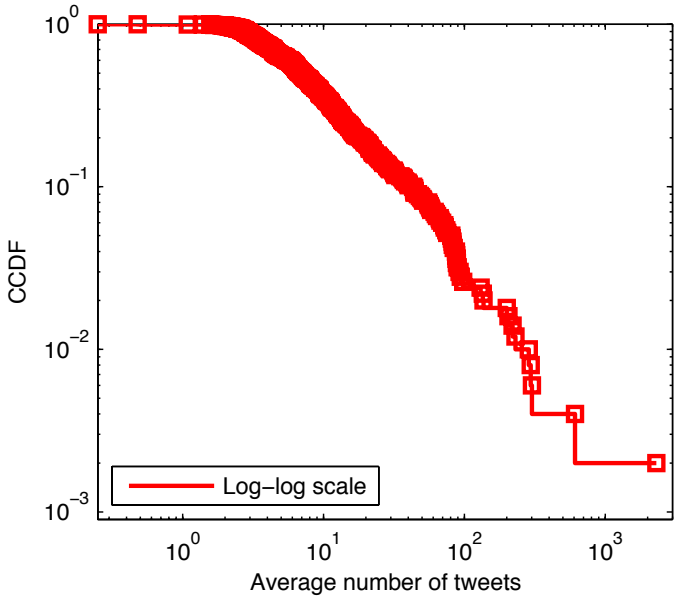


FIGURE 2.2.1: CCDF of the average number of tweets for the S&P 500 stocks.

number of tweets for the stocks is in a wide range, varying from only a few tweets to above 2,000 tweets per day.

We use the daily number of tweets for S&P 500 stocks as the Twitter predictor at stock market level, use the daily number of tweets for each sector as the Twitter predictor at sector level. Table 2.2.1 summarizes the number of companies in each sector and average daily number of tweets we collected for each sector. We notice that Information Technology is the sector have largest average number of daily tweets, which is around 40% of total number of tweets. Financials is the second largest sector in term of both number of companies and average number of daily tweets. At the company stock level, we use the daily number of tweets that mention each company's stock in S&P 500 as the Twitter predictor.

2.2.3 Data normalization

To provide a common scale for comparison of our predictors and stock market indicators, each time series is normalized by its average in the past n trading days. For example, for a dataset X, the normalized time series of x_i in X, denoted as $N(x_i)$, is defined as:

$$N(x_i) = \frac{nx_i}{\sum_{j=i-n}^{i-1} x_j} \tag{2.2.1}$$

In this dissertation, we set n to 70, which is approximately three months of trading days.

2.3 Correlating Number of Daily Tweets with Stock Trading Volume

In this section, we investigate the correlation between number of daily tweets and stock trading volume at each of the three aforementioned levels. The data collected from June 4, 2012 to May 31, 2013, including 240 trading days of are used for correlation analysis.

2.3.1 Stock market level

At the stock market level, we evaluate the correlation between the number of daily tweets and stock market trading volume for S&P 500 introduced in Section 3.2.1. We find that S&P 500 number of daily tweets is positively correlated with daily trading volume with correlation coefficient $r = 0.3$.

TABLE 2.3.1: Correlation coefficient at sector level: correlation between the daily trading volume and the number of daily tweets for each GICS sector.

GICS sector	r
Information Technology	0.31
Financials	0.48
Consumer Discretionary	0.27
Consumer Staples	0.19
Industrials	0.34
Energy	0.37
Health Care	0.13
Materials	0.39
Telecomm Services	0.17
Utilities	0.35

2.3.2 Sector level

At the sector level, we evaluate the correlation between the number of daily tweets and the daily trading volume for each sector. Table 2.3.1 summarizes the results. Six out of ten sectors, Energy, Materials, Industrials, Financials, Information Technology and Utilities sectors have correlation coefficients $r > 0.3$, indicating a significant correlation between the number of daily tweets and the daily trading volume. In particular, financials sector, which is the second largest sector, has a correlation coefficient of 0.48.

2.3.3 Company stock level

At the company stock level, we investigate stocks in S&P 500 that have significant number of daily tweets. Specifically, we investigate 78 out of 500 stocks that have average number of daily tweets larger than 25. Again, we evaluate the correlation between the number of daily tweets and stock trading volume introduced in Section 3.2.1.

Fig. 2.3.1 plots the CDF (cumulative distribution function) of correlation coefficients for 78 stocks. We observe that the number of daily tweets has strong correlation with stock trading volume at company stock level with median of correlation coefficient is 0.53.

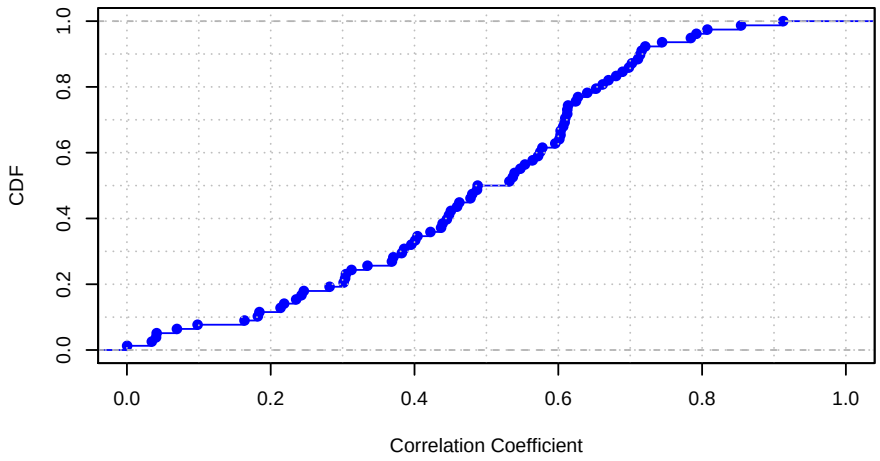


FIGURE 2.3.1: CDF of correlation coefficient for individual stocks

2.4 Predicting Stock Trading Volume Class Using Twitter Data

2.4.1 Trading volume classification

After establishing that the number of daily tweets is correlated with the stock trading volume, we are interested in finding out whether and how well the stock trading volume can be predicted using Twitter data. Instead of predicting the exact value of trading volume, we classify trading volume into three classes: (i) low volume,

(ii) normal volume, and (iii) high volume. Consider a stock. Let $\{Y_t\}$ denote the time series of stock trading volume ratio, where Y_t is the trading volume on day t normalized by the average stock trading volume in the past 70 days. We classify the trading volume ratio on day t into low trading volume C_1 if $Y_t < 0.8$, normal trading volume C_2 if $Y_t \in [0.8, 1.2]$ or high trading volume C_3 if $Y_t > 1.2$.

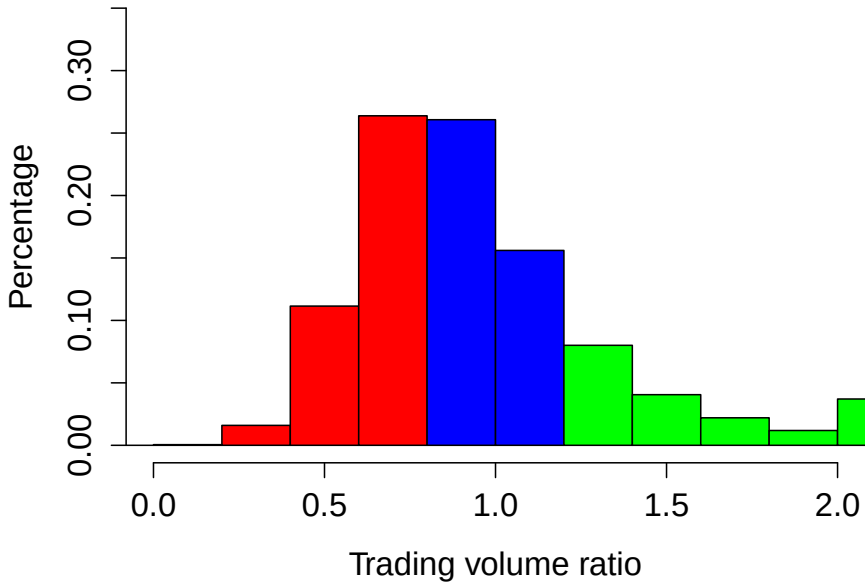


FIGURE 2.4.1: Distribution of stock trading volume ratio

Fig. 2.4.1 plots the distribution of trading volume ratio for 78 stocks that have average daily number of tweets larger than 25 between February 16 , 2012 to May 31, 2013 with total number of 18564 samples. We can see that around 40% of samples are considered as low volume (bars in red), around 40% of samples as normal volume (bars in blue) and around 20% of samples as high volume (bars in green).

2.4.2 Possible Twitter and stock predictors

We now investigate possible predictors to predict stock trading volume. Specifically, we consider the following four predictors: (i) lag 1 tweets ratio, (ii) before-market tweets ratio, (iii) lag 1 trading volume ratio and (iv) interday open close price change rate (short for ioc price change rate in the rest of the dissertation). Consider a stock. We use p^c_{t-1} and p^o_t to denote the daily closing price on day $t-1$ and daily open price on on day t, respectively. The ioc price change rate between day $t-1$ and day t is calculated as the absolute value of the relative price change rate between open price on day t and closing price on day $t-1$, i.e., $(p^o_t - p^c_{t-1})/p^c_{t-1}$. Intuitively, the stock trading volume for a stock may increase significantly when ioc price change is high. On the other hand, if more people post tweets concerning a particular stock before market open, it may indicate people show particular interests in this company's stock and would to trade it.

Consider a stock. Let $\{T_t\}$ denote the time series of tweets ratio, where T_t is the number of tweets on day t normalized by the number of tweets in the past 70 days. Similarly, let $\{T^B_t\}$ denote the time series of before-market tweets ratio, where T^B_t is the number of tweets on day t between 12:00 am to 9:00 am normalized by the number of tweets between 12:00 am to 9:00 am in the past 70 days. Last, let $\{O_t\}$ denote the time series of ioc price change rate, where O_t is the ioc price change rate on day t normalized by the after market price change rate in the past 70 days.

2.4.3 Linear regression

After discuss the possible predictors, we are interested in finding out whether and how well the stock trading volume can be predicted using Twitter data. To answer

this question, we apply a linear regression with exogenous input model using Twitter and stock market predictors as independent variables.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 T_{t-1} + \beta_3 O_t + \beta_4 T_t^B + \varepsilon_t, \tag{2.4.1}$$

where Y_t, O_t, T_t^B represent the stock trading volume ratio, ioc price change rate, and before market tweets ratio on day t, respectively. T_{t-1} and Y_{t-1} represent the tweets ratio and stock trading volume ratio on day $t-1$, respectively. β_0,..., β_4 are regression coefficients need to be determined, and ε_t is a random error term for day t. For each stock, we use the whole 240 days' data collected from June 4 , 2012 to May 31, 2013 to train the regression coefficients and build the model. After that, we use all 240 days' data to fit the model. Specifically, for each stock on each day, we get a fitted value $\hat{Y}_t$ compare with the true value Y_t. Instead of comparing them directly, we classify $\hat{Y}_t$ and Y_t into the corresponding classes introduced in Section 2.4.1 (i.e., if $\hat{Y}_t > 1.2$, it will be classified into C_3), and compare the fitted class with the true class.

We first define following metrics for multi-class problems for our study. True positive (TP) is denoted as the ratio of samples labeled as belonging to a class indeed belonging to this class. In contrast, false positive (FP) is denoted as the ratio of samples labeled as belonging to a class does not belong to this class. True negative (TN) is denoted as the ratio of samples not labeled as belonging to a class indeed belonging to this class. In contrast, false negative (FN) is denoted as the ratio of samples not labeled as belonging to a class does not belong to this class. Our fitting analysis considers three metrics for each class of trading volume: precision, recall, and

TABLE 2.4.1: Fitting performance of linear regression model

Class	TP	TP+FP	TP+FN	Precision	Recall	F1
Low volume	3034	4464	7356	68.0%	41.3%	51.3%
Normal volume	5837	11891	7753	49.1%	75.3%	59.4%
High volume	1362	2209	3455	61.7%	39.4%	48.1%

F_1 score. The precision of a class measures the ratio of samples labeled as belonging to this class does indeed belong to this class (TP) over the total number of samples labeled as belonging to this class (TP + FP), which is the correctness measurement of the model. Recall measures the ratio of TP over the total number of samples belong to this class (TP + FN), which is a completeness measurement of the model. F_1 score is the harmonic mean of precision and recall that evenly weight the recall and precision. Table. 2.4.1 reports the fitting performance using linear regression model. We can see that the precisions achieve 68% and 61.7% for the low volume class and the high volume class, significantly larger than random guess. The F_1 scores achieve 51.3% and 48.1% for low volume class and high volume class, respectively.

2.4.4 Multinomial logistic regression

The multinomial logistic regression model is a regression model that generalizes the linear regression model by allowing for more than two discrete and unordered dependent variables. It is a model that is used to predict the probabilities of the different possible outcomes of a class distributed dependent variable, given a set of independent variables. Multinomial logistic regression allows each class of a dependent variable to be compared to a reference class by providing a number of logistic regression models. As mentioned in Sectionl 2.4.1, the stock trading volume Y_t is classified into three classes C_1, C_2 and C_3, which denoted as low volume, normal volume and high volume,

TABLE 2.4.2: Fitting performance of multinomial logistic regression model

Class	TP	TP+FP	TP+FN	Precision	Recall	F1
Low volume	5101	8143	7356	62.6%	69.3%	**65.8%**
Normal volume	4858	8631	7753	56.3%	62.7%	59.3%
High volume	1276	1790	3455	**71.3%**	36.9%	**48.7%**

respectively. We then build the multinomial logistic regression with two independent binary logistic regression models with one class is selected as the reference class. The multinomial logistic regression is

$$\begin{aligned}\log\frac{\Pr(C_2)}{\Pr(C_1)} &= \beta_{0,2}+\beta_{1,2}Y_{t-1}+\beta_{2,2}T_{t-1}+\beta_{3,2}O_t+\beta_{4,2}T_t^B+\varepsilon_t,\\ \log\frac{\Pr(C_3)}{\Pr(C_1)} &= \beta_{0,3}+\beta_{1,3}Y_{t-1}+\beta_{2,3}T_{t-1}+\beta_{3,3}O_t+\beta_{4,3}T_t^B+\varepsilon_t,\end{aligned} \tag{2.4.2}$$

where class C_1 is selected as the reference class and four independent variables on the right hand side are consistent with the those in Eq 2.4.1. The general multinomial logistic regression model can be written as

$$\log\frac{\Pr(C_v)}{\Pr(C_u)} = \beta_0+\beta_{1,v}X_1+\beta_{2,v}X_2+...+\beta_{p,v}X_p+\varepsilon, \tag{2.4.3}$$

where v is the identified class, u is the reference class, and X_1,...,X_p are denoted as p independent variables.

For each of the 78 stocks, on each day, we select the trading volume class with the largest probability as the fitted class and compare it with the true class. Table. 2.4.2

reports the fitting performance using the multinomial regression model. We can see that the precision of low volume class is 62.6% which is lower than the precision of the low volume class using the linear regression model. However, The F_1 score of the low volume class using the multinomial regression model significantly outperforms that using the linear regression model. Furthermore, for the high volume class, the precision using the multinomial regression model is around 10% larger than that using the linear regression model, and the F_1 score using the multinomial regression model also outperforms that using the linear regression mode. Compare the fitting performance between two models, we find that the multinomial logistic regression model has much better fitting accuracy for the high volume class. For the low volume class, although the precision is lower than that using the linear regression model, the multinomial logistic regression achieves a higher F_1 score. Since we more focus on the high volume precision, we then use the multinomial logistic regression model to predict the stock trading volume in the next section.

2.4.5 Are Twitter predictors useful?

In Section 2.4.4, we find that the multinomial logistic regression model outperforms the linear regression model. In this section, we investigate whether Twitter predictors are useful and can indeed improve the model. To confirm this, we compare the fitting performance for three multinomial logistic regression models, with and without twitter

TABLE 2.4.3: Fitting precision for three models with and without Twitter predictors

Model	Low volume	Normal volume	High volume
2-predictor model	60.7%	53.8%	67.0%
3-predictor model	61.7%	54.7%	70.4%
4-predictor model	62.6%	56.3%	71.3%

predictors.

$$\log \frac{\Pr(C_v)}{\Pr(C_u)} = \beta_0 + \beta_{1,v} Y_{t-1} + \beta_{2,v} O_t + \varepsilon, \tag{2.4.4}$$

$$\log \frac{\Pr(C_v)}{\Pr(C_u)} = \beta_0 + \beta_{1,v} Y_{t-1} + \beta_{2,v} O_t + \beta_{3,v} T_t^B + \varepsilon, \tag{2.4.5}$$

$$\log \frac{\Pr(C_v)}{\Pr(C_u)} = \beta_0 + \beta_{1,v} Y_{t-1} + \beta_{2,v} O_t + \beta_{3,v} T_t^B + \beta_{4,v} T_{t-1} + \varepsilon, \tag{2.4.6}$$

The 2-predictor model in (2.4.4) considers two stock market predictors Y_{t-1}, the trading volume ratio on day $t-1$ and O_t, ioc price change rate on day t. For the 3-predictor model in (2.4.5), we add one twitter predictor T_t^B, the before market tweets ratio on day t. Similarly, for the 4-predictor model in (2.4.6), we add another twitter predictor T_{t-1}, which is the tweets ratio on day $t-1$.

Table 2.4.3 shows the model fitting precision for three models. We can see that, for each class of the trading volume, the precision of the 4-predictor model outperforms that of the 3-predictor model, and the the precision of the 3-predictor model outperforms that of the 2-predictor model, indicate that including Twitter predictors can improve the fitting precision.

Since increasing the number of independent variables can improve the goodness of fit with high probability, to determine which model is the best, we use AICc [4], i.e., Akaike information criterion (AIC) with a correction for finite sample sizes, as a measure of the relative quality of each model. Specifically, for a given statistical

model for m samples, AICc is defined as

$$\begin{aligned} AICc &= AIC + \frac{2k(k+1)}{m-k-1}, \\ AIC &= 2k - 2\ln L, \end{aligned}$$

where k is the number of parameters in the model and L is the maximized value of the likelihood function for the model. A smaller value of AICc indicates the model is preferred. As shown above, AIC is a measure that deals with the trade-off between the goodness of fit of the model and the complexity of the model; AICc enhances AIC by adding greater penalty for extra parameters.

For each of the 78 individual stocks, we calculate the AICc values for the three models described above, denoted as $AICc_2$, $AICc_3$ and $AICc_4$, respectively, where the subscript corresponds to the number of dependent variables in a model. After that, we use paired t-test to pairwise compare the AICc values for the three models. We find that there is strong evidence that $AICc_2 > AICc_3$, $AICc_2 > AICc_4$. However, there is no strong evidence that $AICc_3 < AICc_4$ or $AICc_3 > AICc_4$. The above results indicate that models including Twitter predictors are better than that without Twitter predictors.

We also calculate the residual (mean square error) for the three models, denoted as $Residual_2$, $Residual_3$ and $Residual_4$, respectively, and use paired t-test to pairwise compare the residual values for the three models. We found there is strong evidence that $Residual_2 > Residual_3 > Residual_4$, which further confirms that the model in (2.4.6) with twitter predictors can achieve better fitting accuracy.

TABLE 2.4.4: Prediction performance of multinomial logistic regression model

Class	TP	TP+FP	TP+FN	Precision	Recall	F1
Low volume	1704	2973	2577	**57.3%**	66.1%	61.4%
Normal volume	1728	3260	2963	53.01%	58.3%	55.5%
High volume	424	631	1324	**67.2%**	32.0%	43.4%

2.4.6 Prediction of trading volume class

After establishing that the Twitter predictors are useful for predicting stock trading volume, we are interested in finding out whether and how well the stock trading volume can be predicted using Twitter data. In Section 2.4.4, we confirm that the multinomial logistic regression model is a better choice to predict the stock trading volume compare with the linear regression model. We then use the model in (2.4.2) as our prediction model, with two Twitter predictors and two stock predictors.

In our experiments, for each of the 78 stock, we use the first 150 days of data as training set to build a prediction model, and predict the trading volume for the 151th day. We then use the first 151th days of data as training set to build a prediction model, and predict the results for the 152th day. This process continues for the rest of the data set (which contains data for 88 days).

Table 2.4.4 shows the prediction results for three classes. We mainly focus on the performance of the high trading volume and the low trading volume. We can see that it achieves 57.32% precision to predict low volume with average of 33.8 low volume signals per day. Moreover, it achieves 67.2% precision to predict high volume with average of 7.2 high volume signals per day.

2.5 Summary

In this chapter, we have investigated whether the daily number of tweets that mention S&P 500 stocks is correlated with the stock trading volume. Through correlation analysis, we find correlation between daily number of tweets that mention S&P 500 stocks and stock trading volume at three different levels, from the stock market to industry sector and individual company stocks. Furthermore, we develop two models to predict individual stock trading volume into three categories: low, normal and high. We show that the multinomial logistic regression model outperforms the linear regression model, and it is indeed beneficial to add Twitter data into the prediction models. For the 78 individual stocks that have significant number of daily tweets, the multinomial logistic regression model achieves 57.3% precision for predicting low trading volume and 67.2% precision for predicting high volume.

Chapter 3

Twitter Volume Spikes: Analysis and Application in Stock Trading

3.1 Introduction

Twitter has rapidly gained popularity since its creation and various topics have discussed in Twitter. Stock is a popular topic in Twitter. Many traders, investors, financial analysts and news agencies post tweets about various stocks on a daily basis. The number of tweets concerning a stock varies over days, and sometimes exhibits a significant spike, indicating a sudden increase of interests in the stock. In this chapter, motivated by the observation of Twitter volume spikes, we aim to answer the following questions: (1) When do Twitter volume spikes occur? Are they surprises or expected? What are the potential causes of Twitter volume spikes? and (2) Are Twitter volume spikes useful for stock trading?

In this chapter, we make the following main contributions:

- We find that Twitter volume spikes often happen around earnings dates. Specifically, 46.4% of Twitter volume spikes fall into category. By comparing the implied volatility of a stock before and after a Twitter volume spike, we show that many Twitter volume spikes might be related to pre-scheduled events, and hence are expected to market participants. Furthermore, through correlation analysis, we investigate five possible causes of Twitter volume spikes including stock breakout points, large stock price fluctuation within a day and between two consecutive days, earnings days and high implied volatility. Our results show that only the last two factors show significant correlation with Twitter volume spikes.

- We develop a Bayesian classifier that uses Twitter volume spikes to assist stock trading, and show that it can provide substantial profit. We further develop an enhanced strategy that combines the Bayesian classifier and a stock bottom picking method, and demonstrate that it can achieve significant gain in a short amount of time. Simulation over a half year's stock market data indicates that it achieves on average 8.6% gain in 27 trading days and 15.0% gain in 55 trading days. Statistical tests show that the gain is statistically significant, and the enhanced strategy significantly outperforms the strategy that only uses the Bayesian classifier as well as a bottom picking method that uses trading volume spikes.

The rest of the chapter is organized as follows. Section 3.2 describes data sets and defines Twitter volume spike. Section 3.3 presents the analysis of Twitter volume spikes and possible causes of these spikes. Section 3.4 presents the trading strategies and their performance. Last, Section 3.5 summarizes this chapter.

3.2 Data Collection

3.2.1 Stock market data and Twitter data

In this Chapter, we use stock market data from February 16, 2012 to May 31, 2013 obtained from Yahoo! Finance [70] for the 500 companies in the S&P 500. We use Twitter data collected from Twitter streaming API [67] durning the same period. The datasets for both stock market data and Twitter data are consistent with those datasets used in Chapter 2.

3.2.2 Twitter Volume spike

Consider the tweets concerning a stock. We say the number of tweets on a day is a spike if it is at least K times the average number of tweets in the past N days, $K > 1$. In this dissertation, we set N to 70, i.e., approximately three months of stock market trading days, and set K to 2, 3 or 4. The results presented in the dissertation use $K = 3$; results when $K = 2$ or 4 show similar trends. In this section, we only consider the stocks with average daily number of tweets larger than 10. There are 168 such stocks in S&P 500.

3.3 Twitter Volume Spike Analysis

In this section, we investigate when Twitter volume spikes occur, whether they are surprises or not, and the potential causes of Twitter volume spikes.

3.3.1 When do Twitter volume spikes occur?

We expect that the number of tweets concerning a stock increases sharply when people show particular interests in the stock. One such occasion is company earnings dates, when a company releases earnings reports to inform public their performance during the past time period (most companies release an earnings report each quarter of a fiscal year). People may show particular interests in a company's stock when the company is going to report earnings. In the following, we investigate whether the number of tweets for a stock spikes around the earnings dates.

Suppose that a company's earnings date is day t. We investigate whether the number of tweets on the company's stock spikes around t, in particular, on days $t-1$, t and $t+1$. In our data collection period, there are 509 earnings days for the stocks that we consider. We find 79.2% of them are surrounded by a Twitter volume spike, confirming our intuition that people indeed tweet more about a stock around its earnings dates. Fig. 3.3.1 plots the histogram of the time difference (in days) from an earnings day to the closest day that has a Twitter volume spike, where a negative value corresponds to the time difference to the closest Twitter volume spike in the past. We see that most of the time, the time difference is either 0 (i.e., they are on the same day), 1 (i.e., Twitter volume spike happens on the next day), or -1 (i.e., Twitter volume spike happens on the previous day).

In addition, we mark the Twitter volume spikes that coincide with earnings days, specifically, the spikes that happen within one day (earlier or later) of the earnings days, and find that 46.4% of the Twitter volume spikes fall into this category. This indicates that a significant fraction of Twitter volume spikes happen around earnings days.

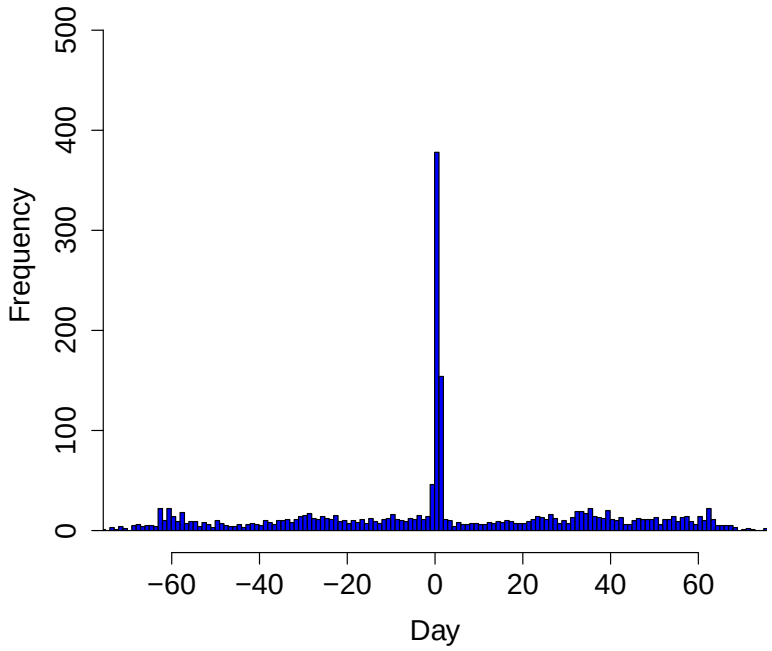


FIGURE 3.3.1: Time difference (in days) from an earnings day to the closest day that has a Twitter volume spike. A negative value corresponds to the the time difference to the closest Twitter volume spike in the past.

3.3.2 Are Twitter volume spikes expected?

Twitter volume spikes close to earnings days are likely due to the earnings days. Since earnings days are public information that people know beforehand, these Twitter volume spikes are no surprises. Certain other scheduled events (e.g., a financial meeting) can also cause Twitter spikes. It is, however, difficult to enumerate such events one by one. On the other hand, we conjecture option implied volatility can be used as an indicator to determine whether a Twitter volume spike is expected or not, that is, whether it is related to a scheduled event. Specifically, we regard a Twitter volume spike as expected when the implied volatility is larger than usual before the

spike happens, and returns back to the usual status after the spike happens. The rationale is as follows. Option implied volatility of a stock indicates how volatile the stock is expected to be based on option prices of the stock. In other words, it indicates how uncertain people feel about the stock. Consider a scheduled event. People anticipate the event, but do not know its impact, hence feel more uncertain about the stock, manifested by the higher implied volatility. Once the event happens, uncertainty reduces, and hence implied volatility returns back to the normal status. When a Twitter volume spike happens between an increased and back-to-normal implied volatility, it is likely related to the anticipated event, and hence is an expected spike. An earnings day described in the previous section is one special case of such expected events. In the following, we obtain the implied volatility of a stock on a day as the weighted average of the implied volatilities of all the options of the stock, where the implied volatility of an option is derived using the Black-Scholes model [12] and the weight for an option is its trading volume on that day.

To shed lights on whether Twitter volume spikes are expected or not, we compare the implied volatility of a stock before and after a Twitter volume spike occurs. Specifically, assume that for a stock, a Twitter volume spike happens on day t. Then we calculate the implied volatility of the stock from day $t-10$ to $t+10$. We consider both short-term options, i.e., those that will expire in 30 days after t, and longer-term options, i.e., those that will expire in 30 to 60 days after t. Fig. 3.3.2 plots the average implied volatility for the ten days before and after a Twitter volume spike, where the index of the days is from -10 to 10, relative to when a Twitter volume spike happens, and the average is obtained considering all the Twitter volume spikes (there are 1245 Twitter volume spikes for all the stocks when $K = 3$). For short-term options, we indeed observe that the daily average implied volatility increases before

t and decreases after t. For longer-term options, the trend is not clear. This might be because option traders usually use short-term options to bet on short-term events to take advantage of higher leverages of short-term options, and hence the implied volatility considering longer-term options is not sensitive to short-term events. For comparison, we also investigate how implied volatility changes before and after a day that is chosen randomly. Specifically, suppose for a stock, a Twitter volume spike happens on day t, then we randomly choose a day t' and calculate the implied volatility of the stock from day $t' - 10$ to $t' + 10$. Fig. 3.3.2 plots the average implied volatility for the ten days before and after such a randomly chosen day, where the average is obtained over all the randomly chosen days (there are 1245 such days). We see the daily average implied volatility shows no significant difference before and after a day that is chosen randomly.

We next use t-test to further confirm the above results. Suppose that for a stock a Twitter spike happens on day t. We only consider the options that will expire in 30 days after t. Let μ_τ^- denote the mean of the daily implied volatility from day $t - \tau$ to $t - 1$. Let μ_τ^+ denote the mean of the daily implied volatility from day $t + 1$ to $t + \tau$. The null hypothesis is that $\mu_\tau^- \leq \mu_\tau^+$. Table 3.3.1 shows the p-values of the t-tests when varying τ from 5 to 10. The very small p-values indicate that we can reject the null hypothesis, indicating that there is strong evidence that $\mu_\tau^- > \mu_\tau^+$, further confirming that the implied volatility before a Twitter volume spike is usually larger than that after the spike. For comparison, we also show the t-test results when choosing a random day, which exhibit large p-values, indicating no strong evidence that $\mu_\tau^- > \mu_\tau^+$.

The above considers the average behavior of all the Twitter volume spikes. Next we consider individual Twitter volume spikes, and identify the percentage of Twitter

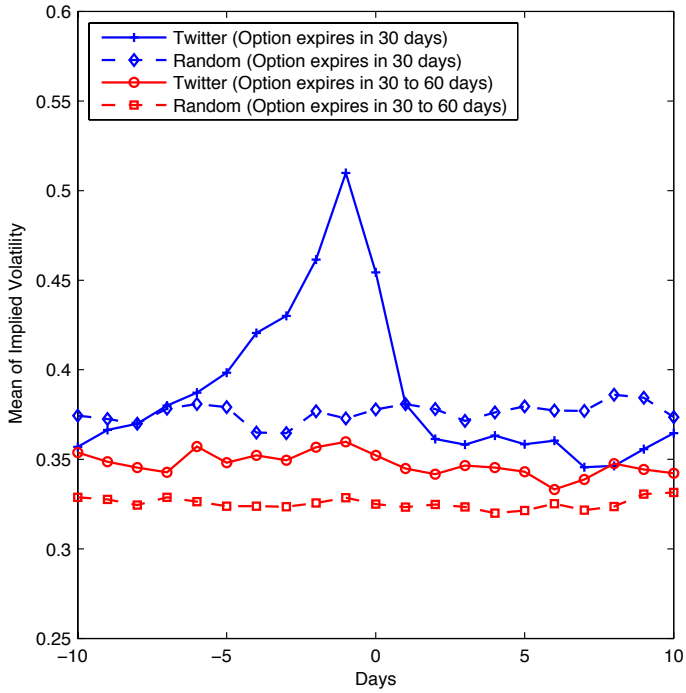


FIGURE 3.3.2: Daily average implied volatility in each of the ten days before and after a Twitter volume spike. Results for randomly chosen days are also plotted in the figure.

volume spikes that are expected. As mentioned earlier, we regard a Twitter volume spike as expected when the implied volatility is larger than usual before the spike, and returns back to the usual status after the spike. To be quantitative, we say a Twitter spike that happens on day t is expected if the implied volatility on $t-1$ is larger than the mean implied volatility in the past N days, i.e., from $t-1-N$ to $t-2$, and the implied volatility on $t+1$ is smaller than the mean of the past N days, i.e., from $t+1-N$ to t, where $N = 70$, i.e., approximately three months of market trading days. We find 37.3% of the Twitter volume spikes satisfy the above condition. Note that this percentage is a very conservative estimate due to the difficulty to quantitatively specify larger and lower than usual. Nonetheless, the result provides a lower bound, indicating that a significant percentage of Twitter volume spikes are expected.

TABLE 3.3.1: p-values of the t-tests for $\mu_\tau^- < \mu_\tau^+$. Only consider options that will expire in 30 days after t.

τ	Twitter volume spike	Random day
5	8.14E-29	0.234
6	1.00E-24	0.510
7	9.71E-23	0.112
8	4.05E-21	0.727
9	6.54E-20	0.763
10	1.52E-17	0.692

3.3.3 Possible causes of Twitter volume spikes

We now investigate potential causes of Twitter volume spikes. Specifically, we consider the following five factors: (i) stock breakout point, (ii) intraday price change rate, (iii) interday price change rate, (iv) earnings day, and (v) stock option implied volatility. In the following, we define the first three factors (the last two factors have been defined earlier), and then calculate the correlation of each of these five factors with Twitter volume spikes.

Consider a stock. We use p_t^c, p_t^h, p_t^l to denote the daily closing price, daily high price, daily low price of the stock on day t, respectively. A stock breakout point is a situation where the price of the stock breaks above a resistance level and rises higher, or breaks below a support level and drops lower. In the following, we say a breakout point happens on day t if the stock closing price is larger or smaller than the closing prices in all of the past N days. We again choose $N = 70$, approximately three months of stock market trading days. The intraday price change rate on day t is calculated as the difference of the daily high price and daily low price, divided by the daily closing price, i.e., $(p_t^h - p_t^l)/p_t^c$. The interday price change rate between day $t-1$ and day t is calculated as the absolute value of the relative price change between

these two days, i.e., $(p_t^c - p_{t-1}^c)/p_{t-1}^c$.

Intuitively, the number of tweets for a stock may increase significantly when a breakout point happens, when it is around an earnings day, or under high intraday price change rate, high interday price change rate, or high implied volatility. In the following, we calculate the correlation coefficient to quantitatively investigate the correlation of Twitter volume spikes and each of the five factors. Consider a stock. Let $\{T_t\}$ denote the time series of Twitter volume spikes, where $T_t = 1$ if there is a Twitter volume spike on day t, and $T_t = 0$ otherwise. Let $\{B_t\}$ denote the time series of stock breakout points, where $B_t = 1$ if there is a stock breakout point on day t, and $B_t = 0$ otherwise. Let $\{C_t\}$ denote the time series of relative intraday price change rate, where C_t is the intraday price change rate on day t normalized by the average intraday price change rate in the past 70 days. Similarly, let $\{D_t\}$ denote the time series of relative interday price change rate, where D_t is the interday price change rate on day t normalized by the average interday price change rate in the past 70 days. Let $\{E_t\}$ denote the time series for earnings days, where $E_t = 0$ by default; while if t is an earnings day, then we set $E_t = 1$, $E_{t-1} = 1$, and $E_{t+1} = 1$ to include one day before and after t. Last, let $\{I_t\}$ denote the time series of relative stock option implied volatility, where I_t is the stock option implied volatility on day t normalized by the average stock option implied volatility in the past 70 days.

We now present lag 1 cross correlation between Twitter volume spikes and each of the five factors, namely, the correlation between T_t and B_{t-1}, the correlation between T_t and C_{t-1}, and so on. The reason for choosing lag 1 is that we are interested in how the value of a factor on the previous day is correlated with the Twitter volume spike on the current day. Fig. 3.3.3 plots the CDF (cumulative distribution function) of the correlations between Twitter volume spikes and each of the five factors over

all the stocks. We observe that Twitter volume spike has the strongest correlation with earnings days (with median of 0.37), which confirms our earlier result that a significant fraction of Twitter volume spikes occurs around earnings days. We also can see that the correlation between Twitter volume spike and implied volatility has a median value of 0.14, much stronger than the correlation with the rest of the three factors.

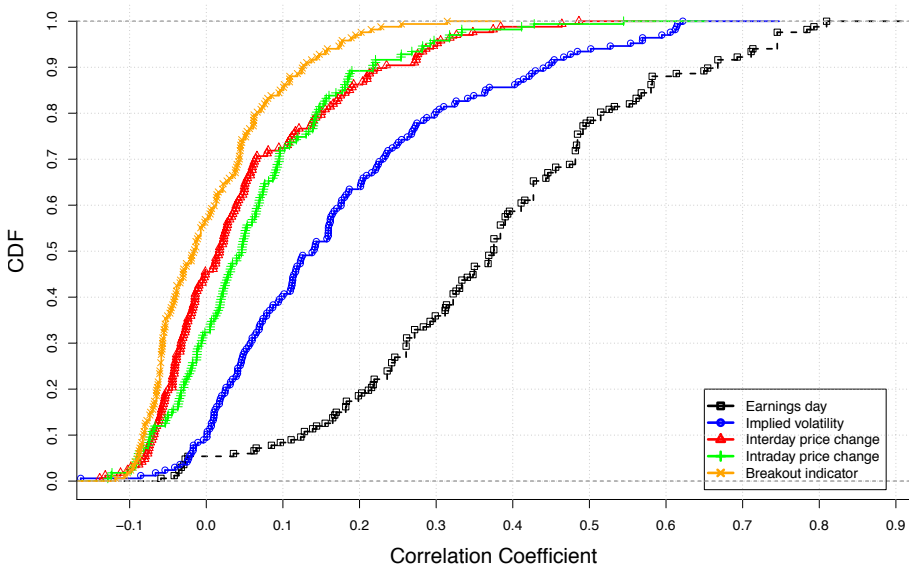


FIGURE 3.3.3: CDF of the lag 1 correlation coefficients between Twitter volume spikes and each of the five factors.

3.4 Application in Stock Trading

After analyzing Twitter volume spikes, a natural question is whether they are useful for stock trading. We develop two trading strategies, both using Twitter volume spikes

as trading signals. We next present these two strategies and their performance. For comparison, we also consider a baseline strategy that purchases a stock on a random day, and a strategy that uses trading volume spikes.

3.4.1 Strategy based on Bayesian classifier

For a stock, after observing a Twitter volume spike, a natural strategy to decide whether to buy the stock or not is as follows. We first calculate the probability that buying the stock can lead to profit after a number of days, and only buy the stock when the probability is sufficiently large. Specifically, we define two types of events, one corresponding to the events that buying the stock leads to profit, and the other corresponding to the opposite, denoted as $\mathcal{G}$ and $\bar{\mathcal{G}}$, respectively. We use a set of features $F_1, \ldots, F_k$ to predict the probability that event $\mathcal{G}$ happens, namely $\Pr(\mathcal{G} \mid F_1, \ldots, F_k)$. Using Bayes rule, we have

$$\Pr(\mathcal{G} \mid F_1, \ldots, F_k) = \frac{\Pr(\mathcal{G}) \Pr(F_1, \ldots, F_k \mid \mathcal{G})}{\Pr(F_1, \ldots, F_k)} \tag{3.4.1}$$

To obtain $\Pr(\mathcal{G} \mid F_1, \ldots, F_k)$, we use a training set to obtain the various probabilities on the right hand side. Obtaining $\Pr(F_1, \ldots, F_k \mid \mathcal{G})$ for large k (i.e., when the number of features is large) requires a large training set. For simplicity, we treat each of the features as independent, so that $\Pr(F_1, \ldots, F_k \mid \mathcal{G}) = \Pi_{i=1}^{k} \Pr(F_i \mid \mathcal{G})$ and obtain $\Pr(F_i \mid \mathcal{G})$ from the training set. Similarly, since $\Pr(F_1, \ldots, F_k) = \Pr(F_1, \ldots, F_k \mid \mathcal{G}) \Pr(\mathcal{G}) + \Pr(F_1, \ldots, F_k \mid \bar{\mathcal{G}}) \Pr(\bar{\mathcal{G}})$, assuming independence, we have $\Pr(F_1, \ldots, F_k) = \Pi_{i=1}^{k} \Pr(F_i \mid \mathcal{G}) \Pr(\mathcal{G}) + \Pi_{i=1}^{k} Pr(F_i \mid \bar{\mathcal{G}}) \Pr(\bar{\mathcal{G}})$, where the various probabilities on the right hand side are obtained from training data.

To evaluate the above strategy, we use the data from February 21, 2012 to October 19, 2012 as training data, and use the data from October 20, 2012 to March 31, 2013 as test data. This results in 573 Twitter volume spikes in the training set, and 672 Twitter volume spikes in the test set. The set of features used in the classifier is a subset of the five factors discussed in Section 3.3.3, excluding implied volatility which requires using option data and hence does not provide a fair comparison with other strategies. Since the stock market closes at 4pm (New York time) each day and the Twitter volume spikes are identified using the number of tweets throughout a day, when observing a Twitter volume spike on day t and we decide to buy the stock, we buy the stock on day $t+1$, using the closing price on day $t+1$. In training, we say buying a stock on day t makes profit if the stock closing price on day $t+10$ is larger than that on day t. In testing, we purchase a stock if the predicted probability, $\Pr(\mathcal{G} \mid F_1, \ldots, F_k)$, is larger than 0.7.

We next report the performance of the above strategy. For a trade that buys stock on day t and sells the stock on day $t+\tau$, we refer to τ as *stock holding period*, and define the *price change rate* as $(p^c_{t+\tau} - p^c_t)/p^c_t$, where p^c_t denotes the closing price on day t. The performance metric we use is average price change rate, defined as the average of the price change rate of all the trades. Fig. 3.4.1 plots the results for three sets of features: breakout point and interday price change rate; breakout point, interday price change rate and earnings days; intraday and interday price change rate. The stock holding period, τ, is varied from 1 to 55 trading days. We can see that the strategy leads to substantial profit. The average price change rate is above 0 from day 10 to day 55 for all the three sets of features, which is also confirmed by t-test (detailed results of the t-test are omitted in the interest of space). In addition, we observe the average price change rate roughly increases over time. Specifically, when

the features are breakout point and interday price change rate, the gain reaches 9.7% when $\tau = 54$ trading days. Fig. 3.4.1 also plots the results of a random strategy. This random strategy differs from our strategy (when the features are breakout point and interday price change rate) in that if our strategy decides to buy stock s on day t, then it decides to buy s on a day that is chosen randomly. We see from the figure that our strategy clearly outperforms the random strategy, which is confirmed by t-test.

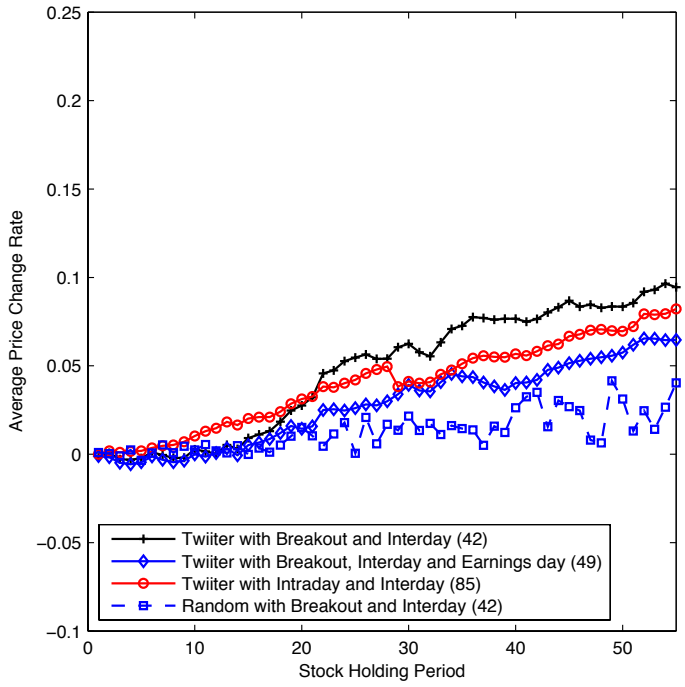


FIGURE 3.4.1: Performance of the strategy based on Bayesian classifier. In the legend of each setting, the number in the parentheses represents the number of trades.

The results of the above simple strategy are encouraging, indicating that Twitter volume spikes are indeed useful in stock trading. On the other hand, the strategy does not consider the trend of a stock. For instance, it may buy a stock when the price of the stock is increasing, which may not lead to profit. In the following, we propose a further enhanced strategy that takes the trends of the stocks into account.

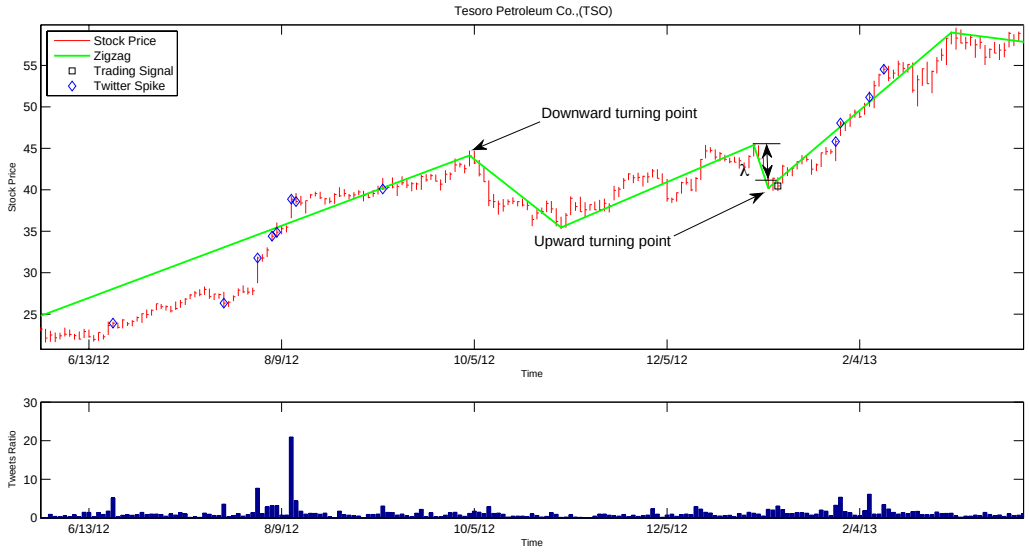


FIGURE 3.4.2: Illustration of the turning points, ZigZag curve and bottom picking method using the price and tweets information of a stock. For the stock, the top figure shows the price over time; the bottom figure shows the tweets ratio, i.e., the number of tweets on a day over the average number of tweets in the past 70 days, over time. A day with tweets ratio above K has a Twitter volume spike.

3.4.2 Enhanced strategy

This strategy uses both Twitter volume spikes and stock turning points. We say that a stock has a turning point on day t when its trend changes on that day. Specifically, a downward turning point indicates that the stock price starts to move downward, and a upward turning point indicates that the stock price starts to move upward, as illustrated in Fig. 3.4.2. We apply a Zigzag based algorithm (based on ZigZag indicator) to identify turning points for a given movement rate, λ, which is defined as the minimum price difference ratio between two adjacent turning points (i.e., the relative difference between two adjacent turning points needs to be at least λ). The stock price turning point identification algorithm for a given λ is described as follows.

- **Stock Price Turning Point Identification Algorithm**

  (1) Start to search from the first point in the data set. Search forward until we find a potential turning point, i.e., one of the two conditions holds: (i) the price increases by at least λ from the start point, or (ii) the price decreases by at least λ from the start point. Continue the search.

  (a) If condition (i) holds (i.e., the price moves upward), update the potential turning point when finding a point that is larger than the previous potential turning point. When finding a point that drops at least λ compared to the current potential turning point, set the current potential turning point to be a downward turning point.

  (b) If condition (ii) holds (i.e., the price moves downward), update the potential turning point when finding a point that is smaller than the previous potential turning point. When finding a point that increases at least λ compared to the current potential turning point, set the current potential turning point to be an upward turning point.

  (2) Start to search from the turning point. If the turning point is a upward turning point, goes to Step (1)(a). If the turning point is a downward turning point, goes to Step (1)(b). Repeat till the end of the data set.

For each stock, we choose the movement rate, λ, based on a stock parameter, β value. The β value of a stock describes the correlated volatility of the stock price in relation to the volatility of the benchmark that the stock is being compared to. In our case, we use S&P 500 index as the benchmark. Specifically, $\beta > 0$ means that the movement of the stock is in the same direction as the movement of the S&P 500 index, and $\beta < 0$ means the opposite; $\beta > 1$ means the movement of the stock is more than

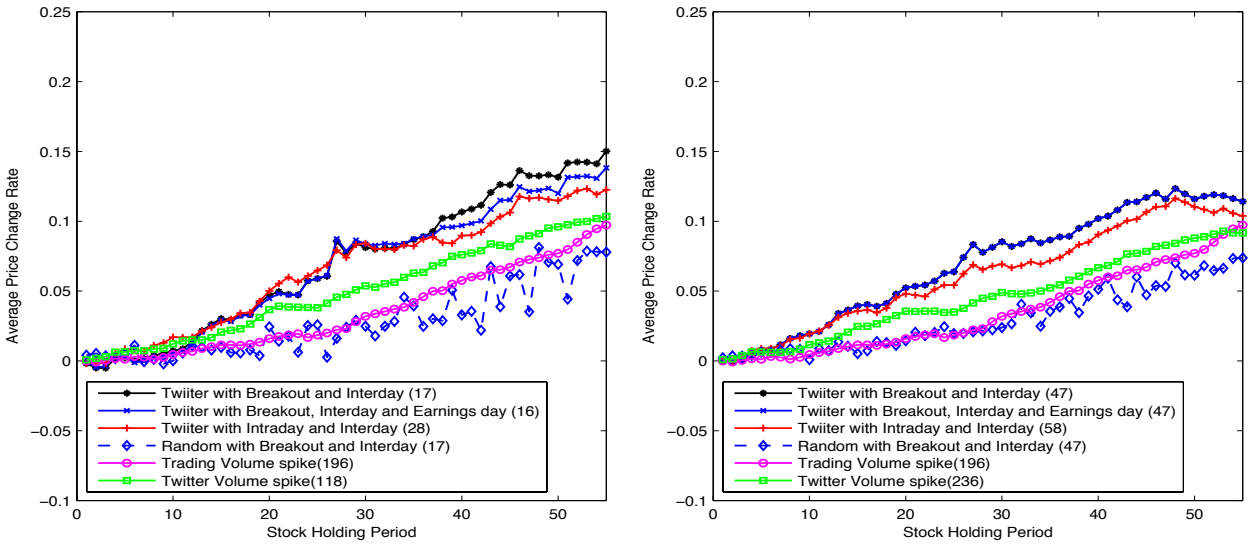
the movement of the S&P 500 index, and $0 < \beta < 1$ means the movement of the stock is less than the movement of the S&P 500 index. For a stock, we use the historical stock closing prices from February 20, 2011 to February 21, 2012 to calculate the stock β value. For stocks with larger β values, we assign a larger movement rate rate. More specifically, we set λ to 10% when $\beta > 1$, and set λ to 7% otherwise. Fig. 3.4.2 illustrates both upward and downward turning points of a stock. The lines connecting two adjacent turning points form the ZigZag curve.

It is clear that using Twitter volume spikes that are close to the bottom of the ZigZag curve (where the stock price is a local minimum) as trading signals can make profit, as illustrated in Fig. 3.4.2. However, when a Twitter volume spike happens, we do not know whether it is close to the bottom because identifying the bottom requires future stock price information. We therefore use the following heuristic method to select Twitter volume spikes that are close to the bottom. First, we only select Twitter volume spikes that happen when the stock price is moving downward, i.e., after a downward turning point. Furthermore, suppose a Twitter volume spike happens on day t, and the previous downward turning point happens on day t'. We only choose the Twitter volume spike when the following two conditions are satisfied: (i) the price has changed by at least λ, i.e., $(p^c_{t'} - p^c_t)/p^c_t \geq \lambda$, where p^c_t is the closing price on day t, and (ii) the closing price on day t is the minimum of the closing prices from t' to t. Since the stock price may fluctuate, we relax the above two conditions by including three earlier days, $t-1$, $t-2$ and $t-3$. That is, if these two conditions are satisfied on one of the four days, from day $t-3$ to day t, then we select the Twitter volume spike as a buy signal. Fig. 3.4.2 shows one such selected Twitter volume spike. Observe that it is indeed close to the bottom. In the following, we refer to the Twitter volume spikes selected using the above bottom picking method as *valid*

Twitter volume spikes.

The enhanced strategy combines the above bottom picking method with the Bayesian classifier described earlier. We again use the data from February 21, 2012 to October 19, 2012 as training data, and the data from October 20, 2012 to March 31, 2013 as test data. This results in 90 valid Twitter volume spikes in the training set, and 118 valid Twitter volume spikes in the test set. To demonstrate that Twitter volume spikes provide valuable information for stock trading, we also compare our strategy with a bottom picking method that is based on stock trading volume spikes, which only differs from our bottom picking method in that it uses stock trading volume spikes, instead of Twitter volume spikes, as trading signals. That is, it uses the same algorithm to identify stock price turning points and the same heuristics to decide whether a day with a stock trading volume spike is close to the bottom of the ZigZag curve of the stock price. To identify stock trading volume spikes, we use the same method for identifying Twitter volume spikes, where we set $N = 70$ and $K = 2$.

We now report the performance of the enhanced strategy. Figures 3.4.3(a) and (b) plot the average price change rate when $K = 3$ and $K = 2$, respectively. The holding period, τ, is again varied from 1 to 55 trading days. The results for three sets of features are plotted in the figure. We observe that the strategy achieves significant gain in a short amount of time. When $K = 3$, the average gain generally increases over time, achieving 8.6% gain when $\tau = 27$, and 15.0% when $\tau = 55$, significantly larger than the gains obtained by the strategy that only uses the Bayesian classifier. When $K = 2$, the gains are also significant (slighter lower than those when $K = 3$), indicating that the strategy is not sensitive to the choice of K. We also observe that our strategy outperforms the random strategy, and the strategy that uses stock trading volume spikes. Last, we observe that the gain when only using valid Twitter



(a) $K = 3$ (b) $K = 2$

FIGURE 3.4.3: Performance of the enhanced strategy. In the legend of each setting, the number in the parentheses represents the number of trades.

volume spikes without using any feature (and hence it does not use the Bayesian classifier) is not as good, indicating it is important to use the bottom picking method along with the Bayesian classifier.

Table 3.4.1 presents the t-test results of the enhanced strategy when $K = 3$. We confirm that there is indeed strong evidence that the profit is positive, and the enhanced strategy outperforms the random strategy as well as the strategy that uses stock trading volume spikes. Specifically, when the features are breakout point and interday price change rate, the profit is positive from day 15 to 55 (the p-values are below 0.02), the profit is larger than that of the random strategy from day 15 to day 40 (the p-values are below 0.1), and is larger than that of the strategy using stock trading volume spikes from day 15 to around day 35 (the p-values are below 0.1). Fig. 3.4.4 plots the number of trades in each month when using the enhanced strategy. We can see that the trades spread in five months' testing period (no trades

in March 2013), instead of in a particular month.

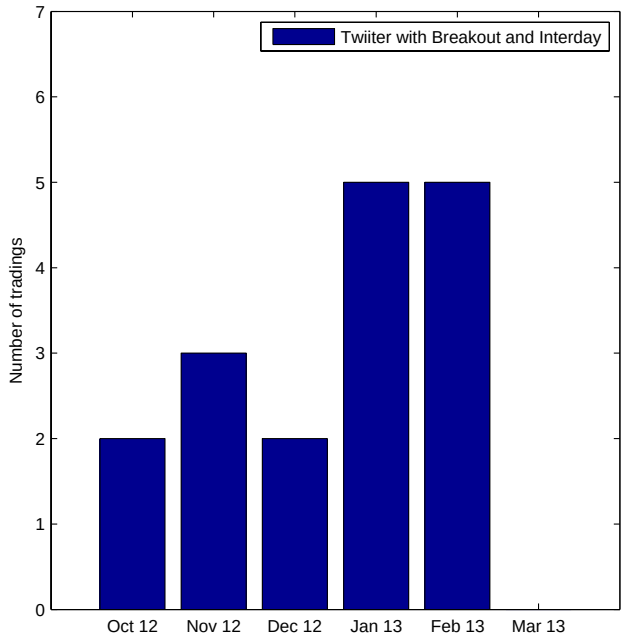


FIGURE 3.4.4: Number of trades in each month when using the enhanced strategy (the features are breakout point and interday price change rate), $K = 3$.

Fig. 3.4.5 plots the fraction of the winning trades (i.e., those that lead to profit) under the enhanced strategy, when the holding period, τ, is varied from 1 to 55 trading days. The results for three sets of features are plotted in the figure. We observe that significant fraction of the trades lead to profit. For instance, when using intraday and interday price change rates as features, 89.3% of the trades lead to profit in 29 days. We also plot the results when only using Twitter volume spikes (i.e., without using any feature) and when using stock trading volume spikes; both show inferior performance compared to the enhanced strategy.

Last, as an example, we present more detailed results using the enhanced strategy and the features are breakout point and interday price change rate. Fig. 3.4.6 plots the profits of the trades in decreasing order (a negative value indicates a loss in money)

TABLE 3.4.1: p-values of the t-tests that compare the profit of the enhanced strategy (for three sets of features) with 0, with the profit using the random strategy, and with the profit using the strategy that is based on stock trading volume spikes.

	Breakout and Interday			Breakout, Interday and Earnings day			Intraday and Interday		
τ	Random	0	Trading vol. spikes	Random	0	Trading vol. spikes	Random	0	Trading vol. spikes
5	0.644	0.372	0.448	0.736	0.351	0.424	0.382	0.083	0.144
10	0.267	0.250	0.410	0.428	0.308	0.466	0.201	0.043	0.122
15	0.084	0.017	0.096	0.311	0.026	0.123	0.039	0.004	0.071
20	0.097	0.012	0.065	0.093	0.020	0.091	0.072	0.000	0.012
25	0.061	0.003	0.029	0.062	0.005	0.036	0.014	0.000	0.002
30	0.037	0.006	0.057	0.055	0.007	0.059	0.012	0.000	0.009
35	0.059	0.014	0.080	0.143	0.019	0.132	0.026	0.001	0.059
40	0.079	0.018	0.158	0.181	0.032	0.219	0.053	0.003	0.158
45	0.107	0.010	0.128	0.093	0.019	0.184	0.059	0.001	0.123
50	0.148	0.017	0.178	0.281	0.030	0.242	0.086	0.002	0.164
55	0.118	0.009	0.188	0.269	0.017	0.253	0.150	0.002	0.263

when the holding period, τ, is 55 trading days. We see 14 out of the 17 trades lead to profit, and the highest gain is 93.4%. Table 3.4.2 shows the detailed information of the 17 trades, including the purchase date, purchase price, tweets ratio (i.e., the ratio of the number of tweets in the Twitter volume spike over the average number of tweets in the past 70 days), the gain when $\tau = 55$, the highest gain and the corresponding τ. We see the highest gains of all the trades are positive. The stock of MHFI is purchased twice, on 2/13/13 and 11/8/12. Fig. 3.4.7 plots the average, maximum and minimum price change rates for each value of τ when varying τ from 1 to 55 trading days. Of all the trades, the largest profit is 95.9%, obtained by purchasing the stock of FSLR (First Solar, Inc.) and $\tau = 52$ trading days. The lowest profit is -15.3% (i.e., loss of 15.3%), caused by purchasing the stock of CF (CF Industries Holdings, Inc.) and $\tau = 38$ trading days. On the other hand, we see from Table 3.4.2 that the highest gain of the trade of CF stock is nonetheless positive.

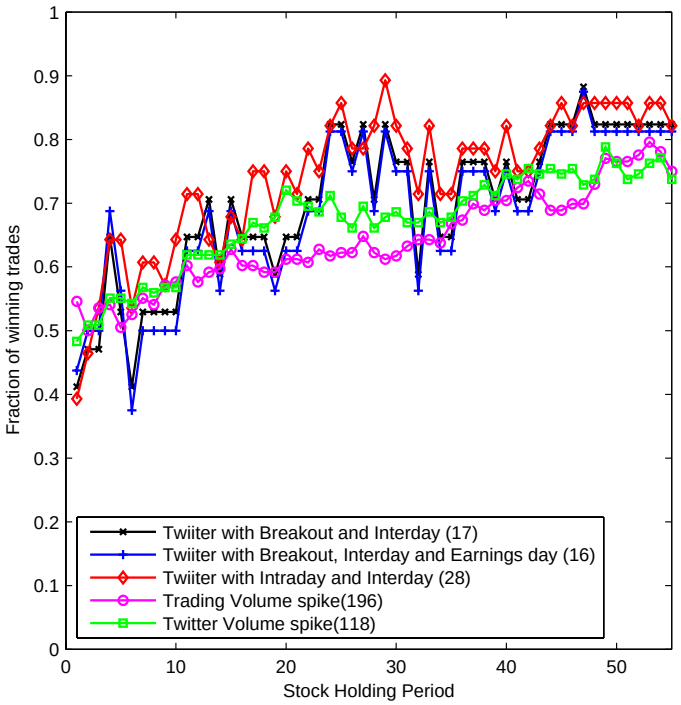


FIGURE 3.4.5: Fraction of the winning trades made using the enhanced strategy, $K = 3$.

3.5 Summary

In this Chapter, we have investigated Twitter volume spikes related to S&P 500 stocks, and whether they are useful for stock trading. Through correlation analysis, we provide insight on when Twitter volume spikes occur and possible causes of these spikes. Moreover, we explore whether these spikes are surprises to market participants by comparing the implied volatility before and after these spikes. Furthermore, we develop a Bayesian classifier that uses the Twitter volume spikes to assist stock trading, and show that it provides slight profit. We further develop an enhanced strategy that combines the Bayesian classifier and a stock bottom picking method, and demonstrate that this strategy can achieve significant gain in a very short amount of time. Simulation over a half year's stock market data indicates that it can achieve

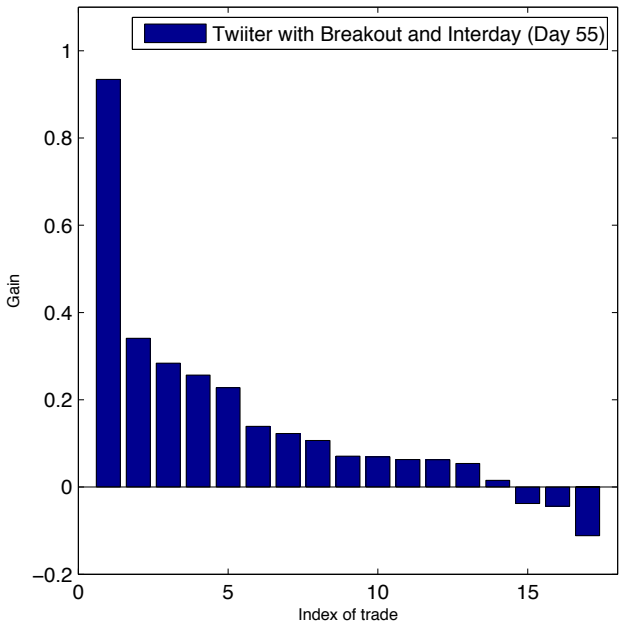


FIGURE 3.4.6: Gains of the trades made using the enhanced strategy, where the features are breakout point and interday price change rate, the holding period τ is 55 trading days, and $K = 3$.

on average 4.8% gain in 15 trading days and 9.5% gain in 33 trading days. Statistical tests show that the gain is statistically significant and the method outperforms the bottom picking method based on trading volume spikes.

TABLE 3.4.2: Summary of the 17 trades made using the enhanced strategy when the features are breakout point and interday price change, $K = 3$.

Stock ticker	Purchase date	Purchase price	Tweets ratio	Gain when $\tau = 55$	Highest Gain (corresponding τ)
FSLR	2/28/13	25.84	3.68	93.42%	95.90% (52)
CELG	11/12/12	75.66	3.10	34.07%	34.07% (55)
TSO	1/10/13	42.85	3.08	28.40%	37.62% (35)
DTV	2/19/13	49.26	3.19	25.66%	25.76% (54)
MHFI	2/13/13	44.33	3.52	22.78%	22.78% (55)
M	12/28/12	37.36	4.76	13.89%	13.89% (55)
MHFI	11/8/12	51.11	5.01	12.27%	12.27% (55)
KSS	1/4/13	42.23	4.77	10.66%	16.79% (48)
FDO	1/4/13	56.65	19.03	7.08%	8.33% (53)
DVN	11/8/12	54.02	4.09	6.90%	6.90% (55)
GME	1/4/13	24.8	4.73	6.29%	8.10% (21)
YUM	2/6/13	62.93	7.72	6.25%	14.32% (35)
EXPE	10/26/12	59.06	6.55	5.38%	9.63% (47)
SHLD	12/11/12	43.5	3.77	1.54%	11.59% (46)
T	10/25/12	34.5	3.44	-3.77%	2.58% (47)
CF	2/21/13	203.93	4.40	-4.39%	2.07% (12)
JDSU	1/31/13	14.51	6.60	-11.16%	6.69% (10)

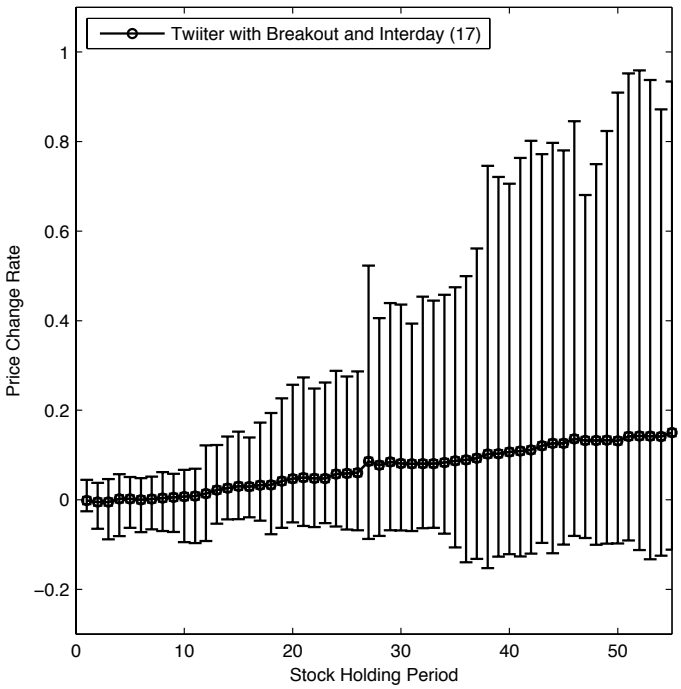


FIGURE 3.4.7: Average, maximum (top bar) and minimum (bottom bar) price change rates of the trades for each value of τ. The results are for the enhanced strategy when the features are breakout point and interday price change, $K = 3$.

Chapter 4

Twitter Volume Spikes and Stock Options Pricing

4.1 Introduction

Twitter has rapidly gained popularity since its creation in March 2006. As of September 2015, it has more than 500 million users, with more than 320 million being active users [66]. The stock market is a popular topic in Twitter. Many traders, investors, financial analysts and news agencies post tweets about the stock market in Twitter, which may be further retweeted. As a result, there can be thousands of tweets each day related to certain stocks. In general, the number of tweets concerning a stock varies over days, and sometimes exhibits a significant spike, indicating a sudden increase of interests in the stock. Since a collection of tweets reflect the collective wisdom of the users who post the tweets, a Twitter volume spike about a stock may contain important information regarding the stock. In this chapter, we investigate

the relationship of Twitter volume spikes and stock options pricing. The reason for focusing on stock options is because they are valuable investment vehicles but are very difficult to understand [47]. Our goal is to investigate whether Twitter volume spikes can shed light on the behavior of stock options pricing, and whether the insights thus obtained can help to assist stock options trading.

A stock option is a financial contract that gives the owner the right, but not the obligation, to buy or sell an underlying asset (stock) at a specified strike price on or before a specified date. Specifically, *call option* gives the owner the right to buy a stock; *put options* give the owner the right to sell a stock. The Black-Scholes model is the most widely used model for stock options pricing. It has led to a boom in options trading ever since it was introduced in 1970's. We start from the underlying assumption of the Black-Scholes model, i.e., stock price follows a geometric Brownian motion and hence stock return follows a lognormal distribution, and investigate when this assumption holds for stocks that have Twitter volume spikes. We then proceed to investigate implied volatility (derived from the Black-Scholes model) as well as the actual volatility around a Twitter volume spike. Our results demonstrate that Twitter volume spikes can be very helpful in understanding stock options pricing. In addition, using Twitter volume spikes, one can design highly profitable options trading strategies. Our main contributions are:

- We find that in a time period with a Twitter volume spike, stock return is less likely to follow a lognormal distribution, indicating that Twitter volume spikes are correlated with extreme changes in stock prices. On the other hand, for a short time period after a Twitter volume spike, the lognormal assumption is likely to hold. In addition, the volatility of a stock is significantly lower after a

Twitter volume spike than that before the spike. We further investigate stock price model selection, and find that a three-parameter model that uses the same drift and different volatilities before and after a Twitter volume spike provides the highest gain in the likelihood value.

- We find a clear pattern in implied volatility (IV) around a Twitter volume spike. Specifically, IV increases sharply before a Twitter volume spike and decreases quickly afterwards. Furthermore, IV of put options tends to be larger than IV of call options. We also find that the volatility around a Twitter volume spike is particularly high. In addition, options may still be overpriced right after a Twitter volume spike. This is particularly true for put options, which confirms that people tend to strongly prefer avoiding losses to acquiring gains [37].

- Based on our findings, we propose a put spread selling strategy for stock options trading. Realistic simulation of a portfolio using one year stock market data demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking account of commissions and ask-bid spread, while S&P 500 only increases 12.8% in the same period.

The rest of the chapter is organized as follows. Section 4.2 describes how we identify Twitter volume spikes. Section 4.3 briefly describes the lognormal stock price model and the Black-Scholes model. Section 4.4 analyzes the relationship between Twitter volume spikes and stock price model. Section 4.5 analyzes the relationship between Twitter volume spikes and stock options pricing. Section 4.6 presents a stock options trading strategy and evaluates its performance. Section 4.7 briefly discusses the choice of threshold for identifying Twitter volume spikes. Last, Section 4.8 summarizes this chapter.

4.2 Methodology

4.2.1 Stock market data and Twitter data

We obtain daily stock market data and stock option data for the 500 stocks in the S&P 500 index. For stock market data, we consider stock daily closing price for each stock. For stock option data, we consider the call and put options of a stock. We only consider short term options that will expire in around 30 days. The Twitter data collection methodology is consistent with what is described in Chapter 2 and Chapter 3. The results reported in this chapter are based on Twitter data collected over one year, from August 1, 2013 to August 6, 2014.

4.2.2 Twitter volume spikes

In this section, we refine our methodology of identifying Twitter volume spikes. Consider a stock. Roughly, a Twitter volume spike happens when the number of tweets related to the stock is significantly larger than usual. Therefore, one way to identify Twitter volume spikes is as follows. We first obtain the number of tweets for the stock on a day and the average number of tweets for the stock in the past N days. Then if the former is significantly larger than the latter, we say there is a Twitter volume spike. The above approach uses the absolute number of tweets to identify Twitter volume spikes, which may not provide robust identification. For instance, it can lead to false Twitter volume spikes when the numbers of tweets for a large number of stocks are inflated (for instance, due to abuse of some users, as we have observed in the collected data). Therefore, for a stock, instead of using the absolute value of the number of tweets, we use the relative value, i.e., the number of tweets

for the stock on a day over the total number of tweets for all S&P 500 stocks on that day, to identify Twitter volume spikes. Specifically, if this relative value is at least K times of the average relative value in the past N days, then we say the stock has a Twitter volume spike. Unless otherwise stated, we use $N = 70$ and $K = 3$ in this dissertation. In Section 4.7, we further investigate the choice of K.

The above definition only considers the number of tweets, while does not consider the users who post the tweets. In our context, a large number of tweets about a stock is only interesting if it indicates that many users show significantly increased interests in the stock. Therefore, we add two additional conditions when identifying Twitter volume spikes. First, the number of unique users has to be sufficiently large. Specifically, we say a stock has a Twitter volume spike on a day only if the number of unique users that post the tweets is larger than a threshold. We choose the threshold to be 10 in this dissertation. Even when the number of unique users is sufficiently large, majority of the tweets can be from a small number of users. To avoid such a scenario, we further require that the tweets have to be from a diverse set of users. Specifically, we define a *user diversity index*, and require that the index to be larger than a threshold. Suppose M unique users tweet about a stock on a day. Let p_i denote the fraction of tweets from user i. Then we define user diversity index as

$$I = \frac{-\sum_{i=1}^{M} p_i \log p_i}{\log M} \tag{4.2.1}$$

where the numerator is the entropy, while the denominator is the maximum value of the entropy (i.e., when each of the M users posts the same number of tweets, i.e., $p_i = p_j, \forall i \neq j$). Therefore, $I \in (0, 1]$. Furthermore, it is easy to see that the value of I is independent of the base of the logarithm by applying change of base in the

logarithm. In this dissertation, we say a stock has a Twitter volume spike on a day only if the user diversity index is above a threshold, chosen as 0.4.

In summary, we use three conditions, one on the number of tweets, one on the number of unique users that post the tweets, and the third on the diversity of the users that post the tweets, when identifying Twitter volume spikes. For the Twitter data that we collected (i.e., tweets that contain S&P 500 stock symbols from August 1, 2013 to August 6, 2014), we find that all the 500 stocks have at least one Twitter volume spike, and there are a total of $3,288$ Twitter volume spikes, which are used in the analysis in the rest of the dissertation.

4.3 Background

The Black-Scholes model assumes that stock price follows a geometric Brownian motion [50]. In the following, we first briefly describe the geometric Brownian motion model, and then describe the Black-Scholes model.

4.3.1 Stock price model

Let S_t denote stock price on day t. Let μ denote the drift rate of the stock, and let σ denote the stock volatility. The most widely used model for stock price [33, 12, 48, 46] is the Geometric Brownian motion model, that is,

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \tag{4.3.1}$$

where W_t is a Brownian motion. On the right hand side of (4.3.1), the first term is used to model deterministic trends, while the second one is used to model unpredictable events. For an arbitrary initial value S_0, the stochastic differential equation (4.3.1) has the analytic solution

$$S_t = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t\right). \tag{4.3.2}$$

Let R_t denote the *log return* (i.e., logarithm of stock return) on day t. Then

$$R_t = \ln \frac{S_t}{S_{t-1}} = \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma\left(W_t - W_{t-1}\right), \tag{4.3.3}$$

where $W_t - W_{t-1}$ is the usual Brownian increment that follows a normal distribution. The above shows that when assuming stock price follows a Geometric Brownian motion, log return follows a normal distribution, or stock return follows a lognormal distribution. Given m samples of log returns, denoted as $\{R_1, \ldots, R_m\}$, the two parameters, μ and σ, can be estimated empirically as

$$\mu = \frac{\sum_{t=1}^{m} R_t}{m}, \quad \sigma = \sqrt{\frac{\sum_{t=1}^{m}(R_t - \overline{R})}{m-1}} \tag{4.3.4}$$

where $\overline{R}$ is the mean of the m samples.

Black-Scholes Model for Stock Option Pricing

A stock option is a financial contract that gives the buyer (owner) the right to buy or sell an underlying asset at a specified price (*strike price*) on or before a specified

date (*expiration date*) [69]. Stock options are in two categories: *call options* and *put options*. A call option of a stock gives the buyer the right to buy the stock at the strike price; a put option gives the buyer the right to sell the stock at the strike price.

The Black-Scholes model [12] is a widely used mathematical model for estimating the price of a stock option. In its basic form, it assumes that the market consists of a risky asset (i.e., a stock) and a riskless asset. The rate of return on the riskless asset is constant, and thus called the risk-free interest rate, denoted as r. The stock does not pay a dividend, and its price follows a Geometric Brownian motion with drift μ and volatility σ. There is no arbitrage opportunity (i.e., there is no way to make a riskless profit). The market is frictionless (i.e., transactions do not incur any fees or costs).

Let t denote time. Let S_t denote the stock price at time t, which is as modeled in (4.3.1). Let $V(S,t)$ be the price of the stock option, which is a function of time t and stock price S. The Black-Scholes equation is a partial differential equation that describes the price of the option over time. Specifically, it is

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0, \tag{4.3.5}$$

where we write $V(S,t)$ simply as V and S_t as S for ease of notation.

The Black-Scholes equation can be used to estimate the price of call and put options. Let T denote its expiration date. Let E denote the strike price of the option. If the option is a call option, it has a payoff of $S_T - E$ if S_T is larger than E. Otherwise, the payoff is zero. That is, the payoff is

$$\max(S_T - E, 0)$$

Using the above condition and the Black-Scholes equation, the price of the call option at time t is

$$S_t N(d_1) - e^{-r(T-t)} E N(d_2), \tag{4.3.6}$$

where $N(d)$ is the cumulative distribution function of the standard normal distribution, and

$$d_1 = \frac{\ln \frac{S_t}{E} + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}, \tag{4.3.7}$$

$$d_2 = \frac{\ln \frac{S_t}{E} + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} = d_1 - \sigma\sqrt{T-t}. \tag{4.3.8}$$

If the option is a put option, it has a payoff of $E - S_T$ if S_T is smaller than E. Otherwise, the payoff is zero. That is, the payoff is

$$\max(E - S_T, 0)$$

Using the above condition and the Black-Scholes equation, the price of the put option at time t is

$$-S_t N(-d_1) + e^{-r(T-t)} E N(-d_2), \tag{4.3.9}$$

where $N(d)$, d_1 and d_2 are as defined earlier.

While the above model assumes no dividend, the case with dividend can also be handled [69]. We address dividend in all the results presented in the dissertation.

4.4 Twitter Volume Spikes and Stock Price Model

While log return is widely assumed to follow normal distribution (see Section 4.3.1), this assumption does not always hold in practice [47]. Specifically, the distribution of log returns can possess much heavier tails than those of normal distribution. In other words, log returns can grow or drop much more sharply than that in normal distribution. Intuitively, sharp increases or decreases in stock returns can trigger more discussions about them, and hence Twitter volume spikes. Therefore, extreme stock prices might be correlated with Twitter volume spikes. In the following, we investigate whether this is indeed the case. After that, we investigate the characteristics of the stock price before and after a Twitter volume spike, and how to choose model parameters in the presence of Twitter volume spikes.

4.4.1 Twitter volume spikes and lognormal assumption

For a stock, consider a time series of log returns over 2τ days around day t, $\mathcal{R}_{t,\tau} = \{R_{t-\tau+1}, \ldots, R_t, \ldots, R_{t+\tau}\}$. In the following, we vary τ from 15 to 150, and identify when $\mathcal{R}_{t,\tau}$ is likely to follow a normal distribution. For this purpose, we consider the log returns of all the S&P 500 stocks from February 21, 2012 to August 1, 2014. For each stock, we random pick a time t and use Shapiro-Wilk test [58] to test whether $\mathcal{R}_{t,\tau}$ follows a normal distribution. Table 4.4.1 shows the percentage of the samples that follow a normal distribution for different values of τ. We can see that as τ increases, the percentage of samples that follow a normal distribution decreases. This indicates that the assumption of normal distribution is more likely to hold for short-term data and is less likely to hold for long-term data. In the rest of the dissertation, we choose $\tau \leq 30$.

TABLE 4.4.1: Percentage of samples that follow a normal distribution.

τ	15	30	50	100	150
Percentage	77.2%	66.2%	53.8%	31.2%	19.6%

TABLE 4.4.2: Percentage of samples that follow a normal distribution for the days around a Twitter volume spike. The results for randomly chosen days are also presented for comparison.

Testing set	$\tau = 15$		$\tau = 30$	
	Twitter vol. spike	Random day	Twitter vol. spike	Random day
$\mathcal{R}_{t,\tau}$	57.4%	76.6%	45.8%	59.4%
$\mathcal{R}^{-}_{t,\tau}$	69.7%	86.0%	60.4%	73.7%
$\mathcal{R}^{+}_{t,\tau}$	83.7%	86.3%	74.6%	76.6%

TABLE 4.4.3: Percentage of samples that follow a normal distribution after excluding days from $t-2$ to $t+3$. The results for randomly chosen days are also presented for comparison.

Testing set	$\tau = 15$		$\tau = 30$	
	Twitter vol. spike	Random day	Twitter vol. spike	Random day
$\mathcal{R}'^{-}_{t,\tau}$	87.9%	88.5%	77.5%	75.9%
$\mathcal{R}'^{+}_{t,\tau}$	88.2%	88.5%	78.1%	77.8%

We next investigate whether extreme stock returns are related to Twitter volume spikes. For this purpose, we consider all the Twitter volume spikes (there are $3,288$ such samples). Suppose for a stock, a Twitter volume spike happens on day t, we then use Shapiro-Wilk test [58] to test whether $\mathcal{R}_{t,\tau}$ follows a normal distribution. Table 4.4.2 shows the percentage of samples that follow a normal distribution, where $\tau = 15$ or 30. For comparison, the results for a day that is chosen randomly are also shown in the table. For fair comparison, the samples of random chosen days are constructed as a one-to-one mapping with those of Twitter volume spikes. Specifically, if for a stock, there is a Twitter volume spike on day t, then we randomly choose a day, t', as a sample that corresponds to the sample for Twitter volume spike. From Table 4.4.2, we see that the log returns around a Twitter volume spike are much less likely to follow a normal distribution than those around a random day.

We next consider the time periods before and after a Twitter volume spike separately. Let $\mathcal{R}^-_{t,\tau}$ denote the series of log returns of τ days, from $t-\tau+1$ to t. We again use Shapiro-Wilk test to test whether $\mathcal{R}^-_{t,\tau}$ follows a normal distribution. Similarly, let $\mathcal{R}^+_{t,\tau}$ denote the set of log returns of τ days, from $t+1$ to $t+\tau$, we test whether it follows a normal distribution. Table 4.4.2 also shows, for each of the two sub-periods, the percentage of samples that follow a normal distribution. We observe that the log returns in the two sub-periods are more likely to follow normal distributions than those in the entire period. Furthermore, the log returns in the latter sub-period (i.e., after the Twitter volume spike, excluding the day with Twitter volume spike) are more likely to follow a normal distribution than those in the former sub-period. This implies that the log returns on the days around a Twitter volume spike, and especially on the day with Twitter volume spike, are more likely to be extreme values. To further confirm this, we remove 6 days, from $t-2$ to $t+3$, in each sample. Specifically,

let $\mathcal{R}'^{-}_{t,\tau}$ denote the set of log returns from $t-\tau+1$ to $t-3$, and let $\mathcal{R}'^{+}_{t,\tau}$ denote the set of log returns from $t+4$ to $t+\tau$. The results are shown in Table 4.4.3. We observe that log returns are indeed more likely to follow a normal distribution after removing these 6 days. In fact, the results are comparable to those when choosing a random day, which further confirms that extreme log returns are correlated with Twitter volume spikes.

4.4.2 Twitter volume spikes and stock price model selection

We have observed that stock price exhibits different behaviors before and after a Twitter volume spike. Specifically, log returns are more likely to follow a normal distribution after a Twitter volume spike. In the following, we first compare the stock volatility in the time periods before and after a Twitter volume spike. The results will provide insights on whether different model parameters are needed for the two time periods. Based on our results in Section 4.4.1, all the results below are restricted to a short time period surrounding a Twitter volume spike. Specifically, suppose that a Twitter volume spike happens on day t. Then we only consider the days in $[t-\tau+1, t+\tau]$, where $\tau \leq 30$. Let σ_τ^- denote the stock volatility derived from the log returns from day $t-\tau+1$ to t. Let σ_τ^+ denote the stock volatility derived from the log returns from day $t+1$ to $t+\tau$. Both σ_τ^- and σ_τ^+ are *empirical* volatility that are obtained using (4.3.4). We use paired t-test to compare σ_τ^- and σ_τ^+ for all $3,288$ Twitter volume spikes. The null hypothesis is $\sigma_\tau^- \leq \sigma_\tau^+$. Table 4.4.4 shows the p-values of the t-tests when varying τ from 15 to 30. The very small p-values indicate that we can reject the null hypothesis, indicating that there is strong evidence that $\sigma_\tau^- > \sigma_\tau^+$. For comparison, we also show the t-test results when choosing a random

TABLE 4.4.4: p-values of the t-tests for $\sigma_\tau^- > \sigma_\tau^+$.

τ	Twitter Volume Spike	Random Day
15	2.5×10^{-71}	0.6
20	7.1×10^{-66}	0.5
25	5.5×10^{-60}	0.5
30	2.1×10^{-46}	0.4

day, which exhibit large p-values, indicating no strong evidence that $\sigma_\tau^- > \sigma_\tau^+$.

The above observation (i.e., $\sigma_\tau^- > \sigma_\tau^+$) indicates that we may need to use different parameters for the two time periods before and after the Twitter volume spike. In the following, we consider three models. The first model uses two parameters for drift and volatility respectively, denoted as $\mu_{2\tau}$ and $\sigma_{2\tau}$, that are estimated from the entire time period (i.e., 2τ days, indicated by the subscripts) using (4.3.4), respectively. The second model estimates three parameters, $\mu_{2\tau}$, σ_τ^- and σ_τ^+, where $\mu_{2\tau}$ is the drift estimated using the entire time period, and σ_τ^-, σ_τ^+ are the volatilities that are estimated using the first τ and last τ days, respectively. The third model uses four parameters, μ_τ^-, μ_τ^+, σ_τ^- and σ_τ^+, where μ_τ^- and σ_τ^- are estimated using the first τ days, and μ_τ^+ and σ_τ^+ are estimated using the last τ days.

To decide which model is the best, we use AICc mentioned in Section 2.4.5, i.e., Akaike information criterion (AIC) with a correction for finite sample sizes, as a measure of the relative quality of each model. For each Twitter volume spike, we calculate the AICc values for the three models described above, denoted as $AICc_2$, $AICc_3$ and $AICc_4$, respectively, where the subscript corresponds to the number of parameters in a model. The value of τ is chosen to be 15, 20, 25 and 30. After that, we use paired t-test to pairwise compare the AICc values for the three models. We find that, for all the settings that we consider, there is strong evidence that $AICc_2 > AICc_3$, $AICc_2 > AICc_4$ and $AICc_4 > AICc_3$. That is, $AICc_2 > AICc_4 >$

$AICc_3$. This is consistent with the earlier results that the volatilities before and after a Twitter volume spike differ significantly, which justifies that they should be estimated separately. On the other hand, the result that the model with three parameters outperforms that with four parameters indicates that it is undesirable to use too many parameters.

Last, we investigate the gain obtained when using a proper model. Let L_2, L_3 and L_4 denote the maximized value of the likelihood function for the three model (the subscript represents the number of parameters in a model). Define the likelihood improvement when using three parameters over using two parameters as $I_3 = L_3/L_2 - 1$. Similarly, define $I_4 = L_4/L_2 - 1$ for the improvement using four parameters over using two parameters. For comparison, we also investigate the case for time period $[t-\tau+1, t+\tau]$ when t is chosen randomly, and denote the likelihood improvements as I_3' and I_4', respectively (in this case, our t-tests also indicate that $AICc_2 > AICc_4 > AICc_3$). We find that the gain when t is a random day is less significant than that when there is a Twitter volume spike on day t. Specifically, we perform t-tests to compare I_3 and I_3', and compare I_4 and I_4'. The null hypotheses are $I_3 \leq I_3'$ and $I_4 \leq I_4'$. Table 4.4.5 shows the p-values of the t-tests when varying τ from 15 to 30. The very small p-values indicate that we can reject the null hypothesis. That is, there is strong evidence that the likelihood improvement corresponding to the case of Twitter volume spike is larger than that of a random day.

In summary, the above results demonstrate that it is important to take Twitter volume spikes into account while studying and modeling stock prices. Specifically, the behavior of stock prices differs significantly before and after a Twitter volume spike: the empirical volatility is lower after a Twitter volume spike, and a three-parameter model that provides separate estimation of the volatilities before and after a Twitter

TABLE 4.4.5: p-values of the t-tests for likelihood improvement.

τ	$I_3 > I_3'$	$I_4 > I_4'$
15	8.6×10^{-13}	5.7×10^{-8}
20	2.5×10^{-12}	1.1×10^{-9}
25	5.1×10^{-10}	1.8×10^{-8}
30	3.9×10^{-8}	2.1×10^{-7}

volume spike provides the highest gain in the likelihood value.

4.5 Twitter Volume Spikes and Stock Options Pricing

Having investigated the relationship of Twitter volume spikes and the lognormal stock price model, we now investigate the relationship of Twitter volume spikes and the Black-Scholes model for stock options pricing. Using the Black-Scholes model, one can derive implied volatility (IV) of an option contract, which is an estimate of the volatility. In the following, we first investigate IV around a Twitter volume spike, and then investigate volatility around a Twitter volume spike.

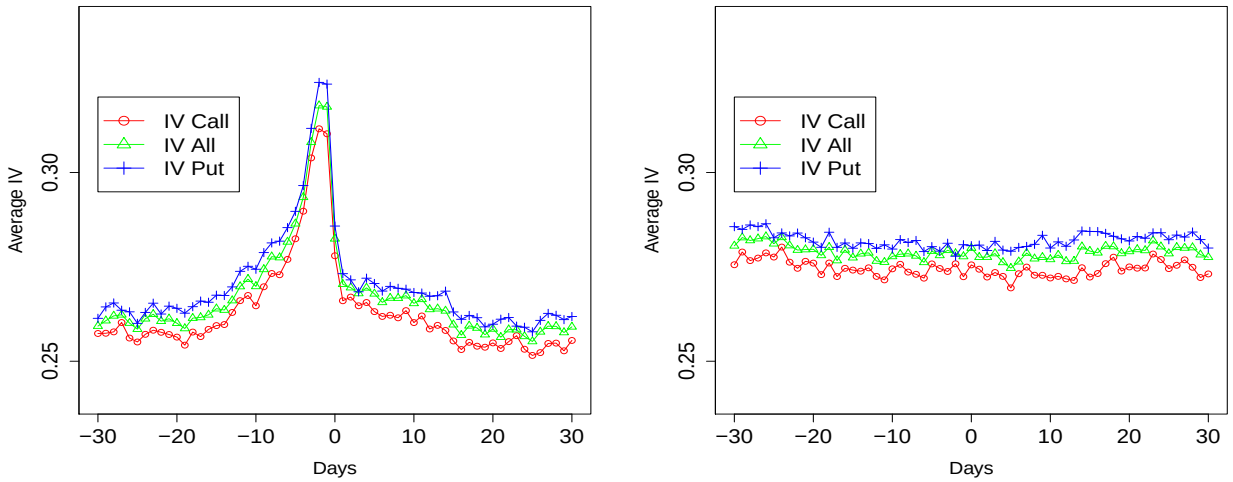
4.5.1 IV around a Twitter volume spike

We only consider short term options that will expire in around 30 days after a Twitter volume spike since long term options are less affected by a Twitter volume spike. Consider a stock. On day t, for a given option price, a given strike price with an expiration date, and the current stock price, we can use (4.3.6) to solve for σ to obtain the IV corresponding to the call option; similarly, we can use (4.3.9) to obtain the IV corresponding to the put option. We obtain IV at the end of a trading day.

The stock price is the daily closing price. The price of an option is taken as the average of the ask and bid prices to take account of ask-bid spread (ask price is the highest price that a buyer is willing to pay for, the bid price is the lowest price for which a seller is willing to sell, and these two prices can be very different).

We next investigate the IV around a Twitter volume spike. For convenience, we represent time as relative to when a Twitter volume spike happens; negative values correspond to days before a Twitter volume spike, while positive values correspond to days after a Twitter volume spike. Suppose that one Twitter volume spike is for a particular stock, and happens on day t_0. We consider all the strikes (that will expire in around 30 days after t_0) for this stock on day t (relative to t_0), and obtain the IVs for the put and call options for each strike on day t. After doing the above for all the Twitter volume spikes, we can obtain the average IV for day t over all the Twitter volume spikes, denoted as $\overline{\sigma_t}$. Specifically, $\overline{\sigma_t}$ is a weighted sum of all the IVs (for each Twitter volume spike, we obtain a set of IVs, one IV for one option), where the weight for an IV is the trading volume of its corresponding option at the end of the trading day. We further obtain two more quantities that are similar to $\overline{\sigma_t}$, denoted as $\overline{\sigma_t^c}$ and $\overline{\sigma_t^p}$, which differ from $\overline{\sigma_t}$ in that $\overline{\sigma_t^c}$ is obtained by only considering call options, while $\overline{\sigma_t^p}$ is obtained by only considering put options.

We next investigate how $\overline{\sigma_t}$, $\overline{\sigma_t^c}$, and $\overline{\sigma_t^p}$ change with t. Fig. 4.5.1 (a) plots these three quantities for $t \in [-30, 30]$. In the figure, for each of these three quantities, the value for day t is an average value that is obtained considering all the instances of Twitter volume spikes, excluding those for which we cannot obtain one of the three quantities (e.g., there may not exist a call or put option with short-term expiration date). We observe that all the three quantities, $\overline{\sigma_t}$, $\overline{\sigma_t^c}$, and $\overline{\sigma_t^p}$, increase sharply before a Twitter volume spike and decrease quickly afterwards. In addition, $\overline{\sigma_t^p}$ is



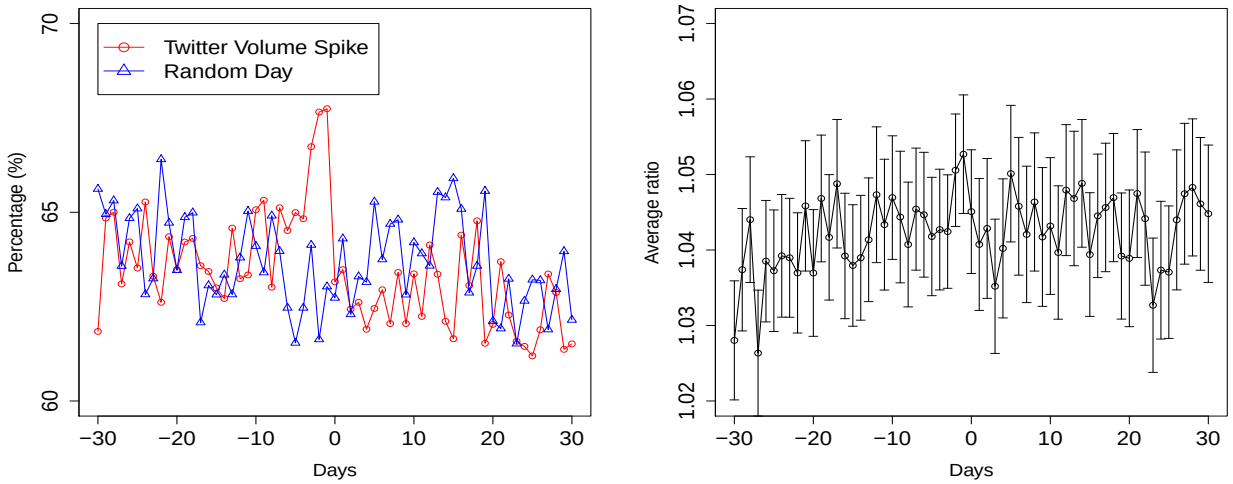
(a) Twitter Volume Spike

(b) Random Day

FIGURE 4.5.1: (a) The average IV for each of the 30 days before and after a Twitter volume spike. Three cases, when only consider call options, only consider put options, and consider all options, are plotted in the figure. (b) The corresponding results for randomly chosen days.

lager than $\overline{\sigma_t^c}$ for all of the 61 days, which indicates that put options may be priced higher compared to call options. We next use t-test to further confirm the above results. The null hypothesis is $\overline{\sigma_t^p} \leq \overline{\sigma_t^c}$ for $t \in [-30, 30]$. For all of the 61 days, 55 days have p-value less than 0.05. The very small p-values on most days indicate that we can reject the null hypothesis, that is, there is strong evidence that $\overline{\sigma_t^p} > \overline{\sigma_t^c}$, further confirming the results we observe from Fig. 4.5.1 (a). For comparison, we also investigate how IV changes before and after a day that is chosen randomly. The results are shown in Fig. 4.5.1 (b). We again observe that $\overline{\sigma_t^p} > \overline{\sigma_t^c}$, which is also confirmed by t-test.

We next further explore the relationship between the IV obtained from put options and the IV obtained from call options. For each Twitter volume spike, for day t, we compare the average IV obtained from put options (again, the average is a weighted



(a) (b)

FIGURE 4.5.2: (a) Percentage that IV obtained from put options is larger than that from call options. (b) Average ratio of IV obtained from put options over IV obtained from call options (with 95% confidence interval).

sum, where the weight for an IV is the trading volume of its corresponding option at the end of the trading day) and that obtained from call options. We then obtain the percentage that the former is larger than the latter considering all the instances of Twitter volume spikes for day t. The results are presented in Fig. 4.5.2(a), $t \in [-30, 30]$. We see that the percentage is above 60% for all 61 days. For the two days immediately before a Twitter volume spike, the percentages are particularly high, and then the percentage drops quickly afterwards. For comparison, Fig. 4.5.2(a) also plots the corresponding results for randomly chosen days, which shows that the percentages are also above 60%. On the other hand, we observe a more significant increase and a more significant decrease in percentage right before and after a Twitter volume spike, compared to the case of random days. The above results again confirm that IV of put options is larger than that of call options. To further illustrate the above points,

we obtain the ratio of the average IV obtained from put options over the average IV obtained from call options for each instance of Twitter volume spike on day t, and then obtain the average ratio over all the instances for day t. Fig. 4.5.2(b) plots the average ratio for each of the 30 days before and after a Twitter volume spike. The 95% confidence intervals are also plotted in the figure. We can see that ratios are above 1.03 for most days, providing further evidence that IV of put options is larger than that of call options.

The above results indicate that the IV is still high right after a Twitter volume spike. A natural question is whether it accurately predicts the actual volatility. In addition, we observe put options are priced higher than call options. A natural question is whether it is rational, or it is due to people's tendency of loss aversion (i.e., people tend to strongly prefer avoiding losses to acquiring gains) [38, 37]. We next answer these two questions by investigating volatility around a Twitter volume spike.

4.5.2 Volatility around a Twitter volume spike

When investigating volatility around a Twitter volume spike, to gain insights, we make a simplifying assumption that the price of a stock follows a Brownian motion (instead of Geometric Brownian motion). That is, we ignore the deterministic term in the right hand side of (4.3.1). This is reasonable since we are only interested in short-term (i.e., within 60 days) behavior. Under this assumption, we have log return on day t as

$$R_t = \ln \frac{S_t}{S_{t-1}} = \sigma \left(W_t - W_{t-1} \right),$$

where W_t is a Brownian motion. From the above, we see that, under the simplifying assumption, R_t/σ follows a standard normal distribution.

We next explore R_t/σ for t around a Twitter volume spike, where t is relative to the day when a Twitter volume happens, $t \in [-29, 30]$. Since we do not know the real σ, we use the IV on day -30 to approximate σ. Specifically, for a stock, the IV on a day is a weighted average considering all the strikes (both call and put options) for the stock (again we only consider strikes that will expire in around 30 days), where the weight is the trading volume of an option at the end of the trading day. We only consider Twitter volume spikes for which we can obtain the IV on day -30. For each such Twitter volume spike, we can obtain one instance of R_t/σ for day $t \in [-29, 30]$. We then use the sample variance to approximate the variance of R_t/σ for $t \in [-29, 30]$. Fig. 4.5.3 plots the results. For comparison, Fig. 4.5.3 also plots the corresponding results for randomly chosen days. We see that for the case of Twitter volume spikes, the variances from day -1 to day 1 are much larger than the corresponding values for the case of randomly chosen days. The difference is most significant on day 0 (the former is 6 times of the latter). On the other hand, for most of the days after 0, i.e., 28 out of 30 days, the variances of the former are lower than those in the latter. The results indicate that the price of a stock is very volatile around a Twitter volume spike (related to this stock), particularly for the days immediately before and after the Twitter volume spike (i.e., for days -1 to $+1$). After that, the volatility is even lower than usual.

The above considers the variance of log returns. We next consider the variance of *cumulative log return*. Consider a stock. Define $R_{t+n,t}$ as the cumulative log return on day $t + n$ relative to day t, $n \geq 1$. Then under the simplifying assumption that

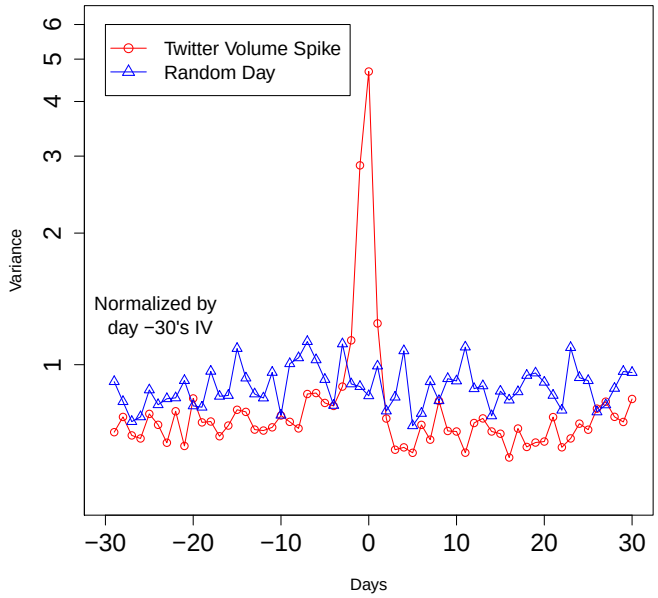


FIGURE 4.5.3: Variance of normalized log returns around a Twitter volume spike.

stock price follows a Brownian motion, we have

$$R_{t+n,t} = \ln \frac{S_{t+n}}{S_t} = \sigma(W_{t+n} - W_t).$$

Therefore, $R_{t+n,t}/\sigma$ follows a normal distribution with variance n.

We now investigate $R_{t+n,t}/\sigma$ around a Twitter volume spike. Again, t is relative to the day when when a Twitter volume happens. We consider $t \in [-30, 30]$. Based on the earlier observation that the variance of log return on a day with a Twitter volume spike is significantly larger than the variances of other days, we divide the time period into two parts, one from day -30 to 0 and the other from day 1 to 30. For the first part, the cumulative log return on day i is $\ln \frac{S_i}{S_{-30}}$, $i \in [-29, 0]$, where S_i is the stock price on day i, and we normalize it by the average IV on day -30.

For the second part, the cumulative log return on day i is $\ln \frac{S_i}{S_1}$, $i \in [2, 30]$ and we normalize it by the average IV on day 1.

Fig. 4.5.4 plots the results for $t \in [-30, 30]$, where the value for t is the average over all the instance of Twitter volume spikes. Specifically, for each t, we have 2955 samples (excluding 333 Twitter volume spikes for which we cannot obtain the IV on day -30 or day 1). For both parts, we use the sample variance of the normalized cumulative log returns to approximate the variance. If the Brownian motion assumption holds, the variance of the normalized cumulative log return will increase linearly with time. For comparison, the corresponding results for randomly chosen days are also plotted in the figure. For the case of randomly chosen days, for both parts (i.e., days $[-30, 0]$ and $[1, 30]$), the variance of cumulative log returns indeed increases approximately linearly with time. For the case of Twitter volume spikes, for both parts, the variance increases linearly with time except for days -2, -1 and 0, which have particularly large variances. We use least squares estimation to estimate the slopes of all the linear curves (for the case of Twitter volume spike, days -2, -1 and 0 are omitted in the estimation). For the case of Twitter volume spikes, the slopes of the two parts are 0.73 and 0.58, respectively, while for the case of random chosen days, the slopes of the two parts are 0.82 and 0.76, respectively. The significantly lower slope of the second part when there are Twitter volume spikes compared to that for randomly chosen days (i.e., 0.58 versus 0.76) indicates that the IV of day 1 (i.e., the day immediately a Twitter volume spike, which is used to normalize $\ln \frac{S_i}{S_1}$, $i \in [2, 30]$) may still be higher than usual, and hence option prices on that day may still be overpriced. This indicates that we can use Twitter volume spike as a trading signal: right after a Twitter volume spike, we can utilize the overpriced options to gain profit, which will be descried in detail in Section 4.6.

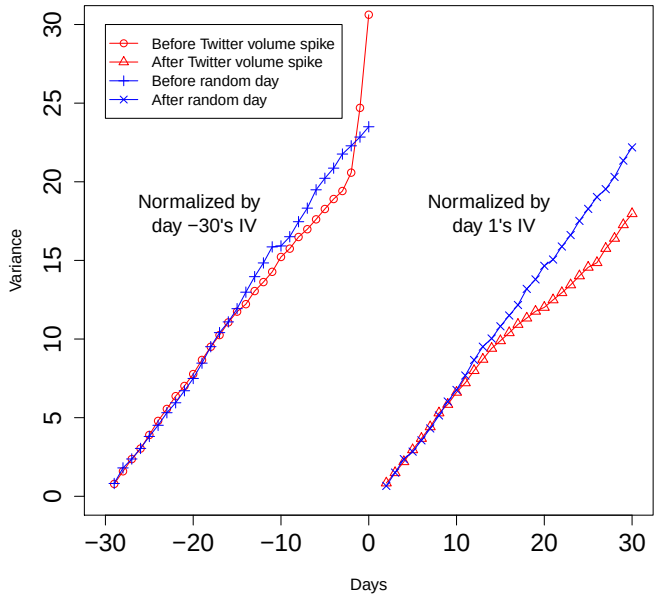


FIGURE 4.5.4: Variance of normalized cumulative log returns around a Twitter volume spike. For comparison, the corresponding results for randomly chosen days are also plotted in the figure.

4.6 Application in Stock Option Trading

Our earlier analysis indicates that put options tends to be priced higher than call options, and option prices may still be overpriced right after a Twitter volume spike. Based on the above results, we conjecture that selling put options right after a Twitter volume spike can be a profitable trading strategy. In the following, we first describe one such strategy and then evaluate its performance.

4.6.1 Put spread selling strategy

Before describing the strategy, we first describe put option selling in more detail. As described earlier, put option is a financial contract between a buyer and seller

of the option. It gives the buyer the right to sell a stock at the strike price on the option expiration day. As an example, suppose that a seller sells a put, which gives a buyer the right to sell 100 shares of the stock of a company, say XYZ, at the strike price of \$80 at expiration (i.e., on the expiration day). To purchase the option, the buyer pays the premium of \$2 per share (premium is paid to the seller of the option and is quoted on a per-share basis). If the stock price is \$82 at expiration, which is higher than the strike price, then the seller can keep the premium, gaining a profit of $2 \times 100 = \$200$. On the other hand, if the stock price drops to \$70 at expiration, then the profit of the buyer is $(80 - 70 - 2) \times 100 = \$800$, while the seller loses \$800. In other words, for a buyer, one of the purposes of buying put option is similar to buying an insurance: it limits the loss of the buyer during unfavorable events with the payment of the premium. For a seller, selling put option can lead to profits through the premium. On the other hand, when the stock price drops significantly, then a seller can lose a substantial amount of money. For instance, in the previous example, if the stock price falls to zero (XYZ bankrupts), then the loss of the seller will be $(80 - 2) \times 100 = \$7800$.

Options spread is widely considered as an option trading strategy to limit the risk. In this dissertation, we consider one type of option spread strategy, called put spread selling. Specifically, the put spread strategy is *bull spread* [47]. It is established with put options by buying a put with a lower strike price and simultaneously selling a put with a higher strike price; the two puts have the same expiration date. This strategy limits the amount of loss. For instance, in the previous example, suppose that a trader buys a put option with the strike price of \$75 at the premium of \$1 per share and sells a put with the strike price of \$80 at the premium of \$2 per share. Then even if the stock price falls to zero, the loss of the trader is limited to

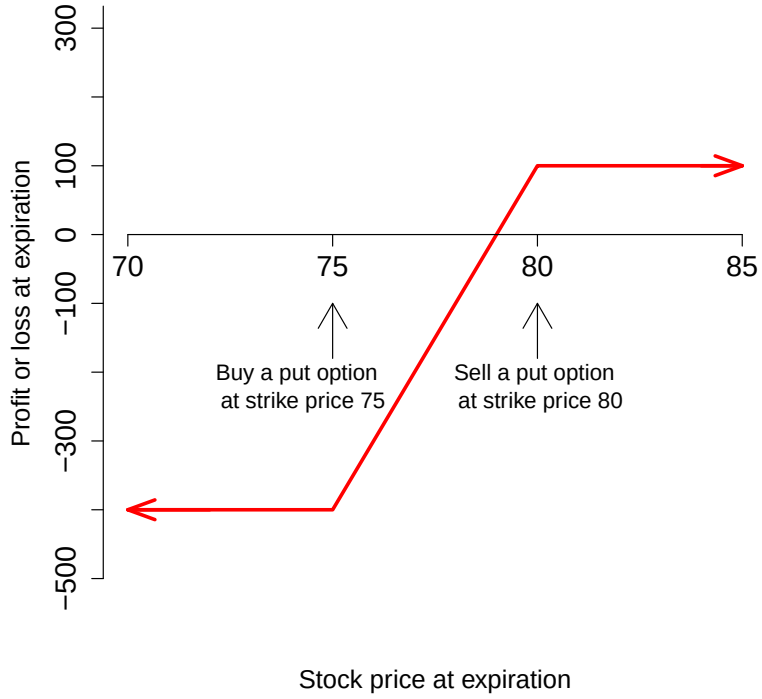


FIGURE 4.6.1: An example illustrating put spread strategy. In the example, the strategy is established by buying a put with the strike price of \$75 at the premium of \$1 per share and selling a put with the strike price of \$80 at the premium of \$2 per share.

$(80 - 2 - 75 + 1) \times 100 = \$400$. Fig. 4.6.1 illustrates the maximum profit and loss (in a negative value) using the above strategy. When the stock price at expiration is no less than \$80, the trader earns a profit of $(2 - 1) \times 100 = \$100$; when the stock at expiration is no more than \$75, the trader has a loss of \$400; and when the stock price at expiration is between \$75 and \$80, the profit of the trader is between $-\$400$ and \$100, and is a linear function of the stock price at expiration.

Based on our observations in earlier sections, we propose the following put spread selling strategy. Suppose that for a stock, a Twitter volume spike happens on day t. Then a trader uses a put spread strategy on a day right after t. Specifically, he will

choose a put spread that will expire in a few weeks after t, and buy and sell puts with δ value in different ranges (δ is used to measure the rate of change of option value with respect to changes in the stock price [69]).

4.6.2 Performance evaluation

We next evaluate the performance of the above strategy. We first consider a simplified simulation scenario, and then consider realistic simulation settings.

Simplified Trading Simulation. In the simplified scenario, we do not consider commission. In addition, the price of an option is set to be the average of the ask and bid prices. The performance metric we use are *premium retention ratio* and *fraction of winning trades*. The premium retention ratio is the amount of profit divided by the amount of premium collected for all traded options. For instance, in the earlier example on bull spread, when the stock price is \$82 at expiration, the premium retention ratio is 1; while when the stock price is \$75 at expiration, the premium retention ratio is $-400/(200-100) = -4$. The fraction of winning trades is defined as the ratio of trades that have positive profit. Table 4.6.1 shows the results, where the trade is on t, $t+1$ or $t+2$ (a Twitter volume spike happens on day t) and the expiration date is four weeks after t. We see that the strategy gains profit in all the settings. Specifically, the average premium retention ratio varies from 31.5% to 59.8%, and the fraction of winning trades varies from 74.2% to 91.4%.

Realistic Trading Simulation. We next evaluate the performance of the put spread strategy through realistic trading simulation. In the simulation, we use a portfolio that can have up to 20 spread positions. Initially, the cash balance is $\$100,000$, the number of open positions is 0, and the number of available positions is 20. After we

TABLE 4.6.1: Performance of the put spread selling strategy in simplified trading simulation.

δ Range		Premium retention ratio			Fraction of winning trades		
Sell	Buy	t	t+1	t+2	t	t+1	t+2
[-0.5,-0.4]	[-0.3,-0.2]	32.6%	31.5%	**33.6%**	74.2%	74.3%	**74.6%**
[-0.4,-0.3]	[-0.2,-0.1]	44.9%	**50.9%**	47.0%	82.5%	**85.6%**	84.4%
[-0.3,-0.2]	[-0.1,0]	48.0%	**59.8%**	54.3%	88.9%	**91.4%**	91.1%

apply put spread strategy for a stock (i.e., sell a put at a high strike price and buy a put at a low strike price), the number of available positions is reduced by one until these options are settled on their expiration day. During the simulation, we try to keep the amount of cash that is allocated to a position to be balanced. Specifically, if c is the current cash balance and n is the number of available positions, then the maximum amount of cash to a position is c/n. For instance, at the beginning, the amount of cash that can be allocated to a position is $100,000/20$. Suppose at a later time, there are already two open positions and the amount of cash is $90,000$. Then the number of available positions becomes 18, and the maximum amount of cash to a position is $90,000/18$. For one position, the number of put spread is $\lfloor c/(nb) \rfloor$, where b is the margin requirement of the put spread (i.e., 100 times the difference of the two strike prices, e.g., in the example in Section 4.6.1, the margin requirement is $(80-75)\times 100 = \$500$). For each put spread, we assume the commission is \$2 (\$1 for selling and \$1 for buying a put). In addition, to be realistic, we take ask-bid spread into account, that is, we buy an option at the ask price and sell an option at the bid price. At any point of time, the number of open positions is no more than 20.

The performance metric is *percentage gain*, that is, the relative difference of the cash balance from the beginning to the end of the simulation. Table 4.6.2 shows the simulation results. This strategy achieves 50.6% rate of return when selling options

TABLE 4.6.2: Performance of the put spread selling strategy in realistic trading simulation.

δ Range		Trading day		
Sell	Buy	t	t+1	t+2
[-0.5,-0.4]	[-0.3,-0.2]	-18.3%	-32.5%	**10.6%**
[-0.4,-0.3]	[-0.2,-0.1]	19.9%	**50.6%**	37.4%
[-0.3,-0.2]	[-0.1,0]	-0.8%	**34.3%**	33.8%

with $\delta \in [-0.4, -0.3]$ and buying options with $\delta \in [-0.2, -0.1]$ on the day following a Twitter volume spike. Although this setting achieves high rate of return, the stock volatilities for this setting are also relatively large, indicating that the trading risk for this setting is large. When selling options with $\delta \in [-0.3, -0.2]$ and buying options with $\delta \in [-0.1, 0]$, which is a lower risk setting, the strategy still achieves 34.3% rate of return. Fig. 4.6.2 plots the simulation result for this setting. The upper figure shows the value of the asset (available cash plus value of the options) on each day, and the lower figure shows the number of open positions on each day. We observe that both quantities change in a stable fashion. Last, of all 180 tradings, only 17 tradings lose money. The fraction of winning trades is 90.6%.

4.7 Choice of Threshold

So far, we have used threshold $K = 3$ when identifying Twitter volume spikes. In this section, we investigate how to choose K. The approach we use is based on the insights on how average IV changes around a Twitter volume spike. Let $\mathcal{D}$ denote the set of Twitter volume spikes that are identified using $K = 3$. Let $\mathcal{D}'$ denote the set of Twitter volume spikes that are identified using $K' \neq K$. It is clear that $\mathcal{D} \subseteq \mathcal{D}'$ when $K' < K$. When $K = 3$, let $\overline{\sigma_t}$ denote the average IV over all the instances of

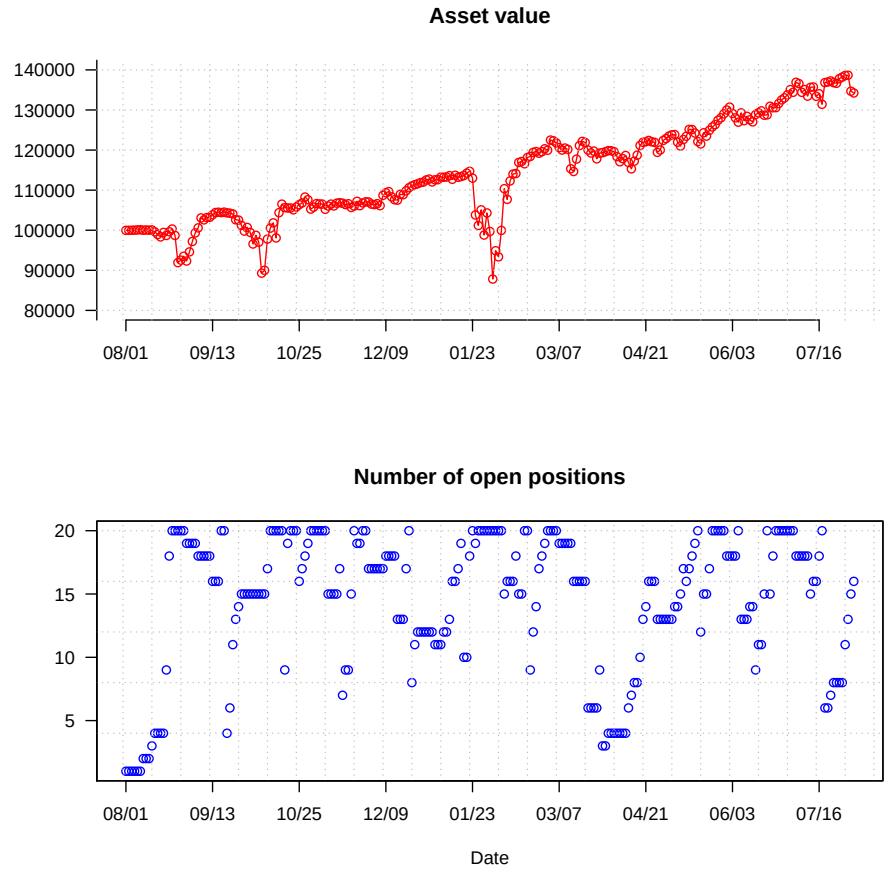


FIGURE 4.6.2: Put spread simulation. The setting is: sell options with $\delta \in [-0.3, -0.2]$ and buy options with $\delta \in [-0.1, 0]$. The upper figure shows the value of the asset (available cash plus value of the options) on each day; the lower figure shows the number of open positions in the portfolio.

Twitter volume spikes as calculated in Section 4.5.1, where t is relative to the day when a Twitter volume spike happens, $t \in [-30, 30]$. For $K' < K$, let $\overline{\sigma'_t}$ denote the average IV over all the instances of Twitter volume spikes in $\mathcal{D}'$, and let $\overline{\sigma''_t}$ denote the average IV over all the instances of Twitter volume spikes in $\mathcal{D}' \setminus \mathcal{D}$, that is, $\overline{\sigma''_t}$ is the average IV from the additional Twitter volume spikes when choosing a smaller K'.

Define the distance between $\overline{\sigma_t}$ and $\overline{\sigma'_t}$ as the normalized Euclidean distance. Sim-

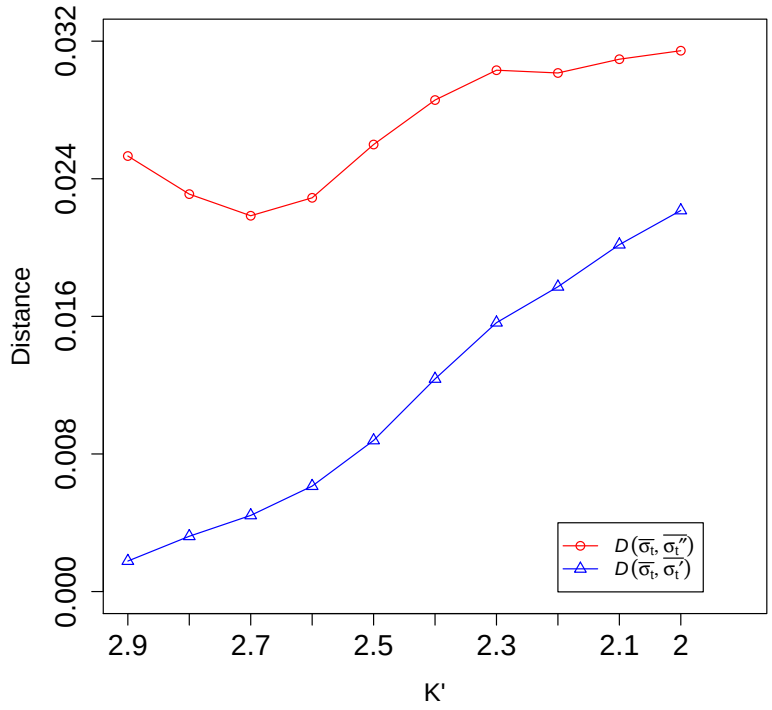


FIGURE 4.7.1: The distance between $\overline{\sigma_t}$ and $\overline{\sigma_t'}$ (the lower curve with triangles) and the distance between $\overline{\sigma_t}$ and $\overline{\sigma_t''}$ (the upper curve with circles) when K' decreases from 2.9 to 2.

ilarly, define the distance between $\overline{\sigma_t}$ and $\overline{\sigma_t''}$. That is,

$$D\left(\overline{\sigma_t}, \overline{\sigma_t'}\right) = \sqrt{\frac{\sum_{t=-30}^{30}\left(\overline{\sigma_t} - \overline{\sigma_t'}\right)^2}{61}}, \quad D\left(\overline{\sigma_t}, \overline{\sigma_t''}\right) = \sqrt{\frac{\sum_{t=-30}^{30}\left(\overline{\sigma_t} - \overline{\sigma_t''}\right)^2}{61}}$$

Fig. 4.7.1 plots the distances defined above when K' decreases from 2.9 to 2. As expected, $D\left(\overline{\sigma_t}, \overline{\sigma_t'}\right)$ increases when K' decreases (i.e., deviates more from 3). The slope of the increase is lower at the beginning and becomes larger afterwards. The distance $D\left(\overline{\sigma_t}, \overline{\sigma_t''}\right)$ is the minimum when $K' = 2.7$. The larger distance when K' is larger than 2.7 is due to a small number of samples in $\mathcal{D}' \setminus \mathcal{D}$. When K' is smaller than 2.7, more Twitter volume spikes are identified; on the other hand, $\overline{\sigma_t''}$ deviate more

from $\overline{\sigma_t}$, leading to larger distances. The above results indicate that the threshold can be chosen from 2.7 to 3, which may achieve similar performance as that when choosing the threshold to 3.

To further confirm this, we use $K = 2.7$ and the same thresholds for the number of unique users and user diversity index to identify Twitter volume spikes. In this case, we identify $4,088$ Twitter volume spikes (24.3% higher than that when using $K = 3$). We then repeat the analysis presented in Section 4.4 to 4.6 using the new set of Twitter volume spikes. Indeed, we find that the observations on stock price, IV and volatility are similar as those when $K = 3$, and the performance of the put spread trading strategy is similar as that when $K = 3$.

4.8 Summary

In this chapter, we have investigated the relationship between Twitter volume spikes and stock options pricing. We started with the underlying assumption of the Black-Scholes model, and investigated when this assumption holds for stocks that have Twitter volume spikes. We next investigated stock volatility around a Twitter volume spike and found that a three-parameter model that uses the same drift and different volatilities before and after a Twitter volume spike provides the highest gain in the likelihood value. We also found a clear pattern in IV around a Twitter volume spike: IV increases sharply before a Twitter volume spike and decreases quickly afterwards. In addition, put options tend to be priced higher than call options. Last, we found that right after a Twitter volume spike, options may still be overpriced. Based on the above findings, we propose a put spread selling strategy. Realistic simulation

using one year stock market data demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking account of commissions and ask-bid spread, while S&P 500 increases 12.8% in the same period.

Chapter 5

Related Work

In this chapter, we briefly review several directions of research that are related to this dissertation.

Twitter analysis and event prediction. The early studies on Twitter have investigated the general characteristics of the Twitter social network (e.g., [34], [42]) and the social interactions within Twitter [32]. Java et al. [34] find that people use Twitter most often to talk about their daily activities and to seek or share information. Krishnamurthy et al. [42] identify distinct classes of Twitter users and their behaviors, and Huberman et al. [32] find that most of the links declared within Twitter are meaningless from an interaction point of view. Weerkamp et al. [18] look at the ways people engage in conversation on Twitter, and found differences between nationalities. Several studies look at the influence of users in Twitter (e.g., [53], [7], [19]). Cha et al. [19] analyze the influence of Twitter users by focusing on users' followers, retweets, and mentions. Romero et al. [53] find that high popularity does not necessarily imply high influence and vice versa. Bakshy et al. [7] find that the largest cascades tend

to be generated by users who have been influential in the past and who have a large number of followers. Later on, several studies try to predict information propagation in social media (e.g., [63], [5], [64]). Artzi et al. [5] propose a model for predicting the likelihood of a response or a retweet on the Twitter network, and Tan et al. [64] study the effects of wording on information propagation. Other studies investigate the public sentiment using Twitter data [54, 30, 10, 41, 3, 23, 35, 11, 55, 51]. They investigate the utility of linguistic features for detecting the sentiment of Twitter messages, and propose approaches for automatically classifying the sentiment of Twitter messages.

In another direction, several studies use tweets to predict real-world events such as earthquakes [56, 21], box-office revenues of movies [6, 25], seasonal influenza [2, 22, 1], sport games, such as NFL [60] and FIFA world cup [20], election [65], the popularity of a news article on Twitter [8], and popular messages in Twitter [31, 15]. Hone et al. [31] propose a method to predict messages which will attract more retweets, and Boyd et al. [15] investigate how authorship, attribution, and communicative fidelity are negotiated in diverse ways by analyzing retweets information.

Stock prediction using Twitter data. Exists studies that are closest to ours are those that relate Twitter to the financial market. Kanungsukkasem et al. [39] propose a method to recognize NASDAQ stock symbols in a stream of tweets. Eduardo et al. [54] report there exists correlation between trading volume and the daily number of tweets for individual company stocks. Bar-Haim et al. [9] predict stock price movement by analyzing tweets to find expert investors and collect experts' opinions. Sprenger et al. [61] find an association between tweet sentiment and stock returns, message volume and trading volume, as well as disagreement and volatility. Schumaker et al. [57] evaluate the sentiment in financial news articles. Several studies use Twitter sentiment data to predict the stock market. Bollen et al. [14] find that

specific public mood states in Twitter are significantly correlated with the Dow Jones Industrial Average (DJIA), and thus can be used to forecast the direction of DJIA changes. Mittal et al. [49] based on [14]'s research apply sentiment analysis and machine learning principles to find the correlation between public mood and stock market movements. Zhang [72] investigates the stock market return using Twitter sentiment data based on three different models. Zhang et al. [74] find that emotional tweet percentage is correlated with DJIA, NASDAQ and S&P 500. Later on, Mao et al. [45] find that Twitter sentiment indicator and the number of tweets that mention financial terms in the previous one to two days can be used to predict the daily market return. Zhang et al. [73] predict financial market movement such as gold price, crude oil price, currency exchange rates and stock market indicators by analyzing Twitter sentiment posts. Si et al. [59] propose a technique to leverage topic based sentiments from Twitter to help predict the stock market. Makrehchi et al. [44] propose an approach that uses event based sentiment tweets to predict the stock market movement, and develop a stock trading strategy that outperforms the baseline.

Our current study differs from all the above in that we focus on Twitter volume spikes and stock market, including both stock pricing and stock options pricing. To the best of our knowledge, this is the first study that investigate how Twitter volume spikes can be used to understand stock market and assist stock trading.

Chapter 6

Conclusion & Future Work

In this dissertation, we have investigated using the tweets concerning S&P 500 stocks to analyze the stock markets and assist stock trading.

In the first part of dissertation, we investigated the correlation between Twitter data and stock trading volume, and predicting stock trading volume using Twitter data. We investigated whether the daily number of tweets that mention S&P 500 stocks is correlated with the stock trading volume, and find correlation at three different levels, from the stock market to industry sector and individual company stocks. We then developed two models, one based on linear regression and the other based on multinomial logistic regression, to predict individual stock trading volume into three categories: low, normal and high. We found that the multinomial logistic regression model outperforms the linear regression model, and it is indeed beneficial to add Twitter data into the prediction models. For the 78 individual stocks that have significant number of daily tweets, the multinomial logistic regression model achieves significant precision for predicting low trading volume and high trading volume.

In the second part of dissertation, we have investigated Twitter volume spikes related to S&P 500 stocks, and whether they are useful for stock trading. Through correlation analysis, we provided insight on when Twitter volume spikes occur and possible causes of these spikes. Moreover, we explored whether these spikes are surprises to market participants by comparing the implied volatility before and after these spikes. After that, we developed two trading strategies that use Twitter volume spikes, one is a basic strategy based on Bayesian classifier and the other is an enhanced strategy that combines the Bayesian classifier and a stock bottom picking method. Simulation over a half year's stock market data demonstrates that both strategies lead to substantial profits, and the enhanced strategy significantly outperforms the basic strategy and a bottom picking method that uses trading volume spikes.

In the third part of dissertation, we have investigated the relationship between Twitter volume spikes and stock options pricing. We started with the underlying assumption of the Black-Scholes model, and investigated when this assumption holds for stocks that have Twitter volume spikes. We next investigated stock volatility around a Twitter volume spike and found that a three-parameter model that uses the same drift and different volatilities before and after a Twitter volume spike provides the highest gain in the likelihood value. We also found a clear pattern in IV around a Twitter volume spike: IV increases sharply before a Twitter volume spike and decreases quickly afterwards. In addition, put options tend to be priced higher than call options. Last, we found that right after a Twitter volume spike, options may still be overpriced. Based on the above findings, we propose a put spread selling strategy. Realistic simulation over seven and half months stock market data to demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking

account of commissions and ask-bid spread, while S&P 500 increases 12.8% in the same period.

As future work, we are pursuing in two directions: (1) looking into the content of tweets to understand their impact on stock pricing and stock options pricing. (2) considering more sophisticated Twitter volume spike metrics, and (3) adding more realistic trading constrains in our stock and options trading strategy.

Bibliography

[1] H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu, and Benyuan Liu. "Predicting Flu Trends using Twitter data". In: *Proc. of Computer Communications Workshops (INFOCOM Workshop)*. ShangHai, China, 2011.

[2] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu Hsin Yu, and Benyuan Liu. "Twitter Improves Seasonal Influenza Prediction". In: *Proc. of Annual International Conference on Health Informatics (HEALTHINF)*. Vilamoura, Algarve, Portugal, 2012.

[3] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment Analysis of Twitter Data". In: *Proc. of the Workshop on Languages in Social Media*. Portland, Oregon, 2011.

[4] H. Akaike. "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723.

[5] Yoav Artzi, Patrick Pantel, and Michael Gamon. "Predicting responses to microblog posts". In: *Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montreal, Canada, 2012.

[6] S. Asur and B.A. Huberman. "Predicting the Future with Social Media". In: *Proc. of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Toronto, Canada, 2010.

[7] Eitan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. "Everyones an influencer: Quantifying influence on Twitter". In: *Proc. of the fourth ACM international conference on Web search and data mining*. Hong Kong, China, 2011.

[8] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. "The Pulse of News in Social Media: Forecasting Popularity". In: *CoRR* abs/1202.0332 (2012).

[9] Roy Bar-Haim, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. "Identifying and Following Expert Investors in Stock Microblogs". In: *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, United Kingdom, 2011.

[10] Luciano Barbosa and Junlan Feng. "Robust Sentiment Detection on Twitter from Biased and Noisy Data". In: *Proc. of the 23rd International Conference on Computational Linguistics*. COLING '10. Beijing, China, 2010.

[11] Albert Bifet and Eibe Frank. "Sentiment Knowledge Discovery in Twitter Streaming Data". In: *Proc. of the 13th International Conference on Discovery Science*. DS'10. Canberra, Australia, 2010.

[12] Fischer Black and Myron Scholes. "The Pricing of Options and Corporate Liabilities." In: *Journal of Political Economy* 81.3 (1973), pp. 637–654.

[13] *Bloomberg Finance*. http://www.bloomberg.com/quote/SPX:IND.

[14] Johan Bollen, Huina Mao, and XiaoJun Zeng. "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1 (2011), pp. 1 –8.

[15] D. Boyd, S. Golder, and G. Lotan. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter". In: *Proc. of 43rd Hawaii International Conference on System Sciences (HICSS)*. Honolulu, Hawaii, 2010.

[16] Kirt C. Butler and S.J. Malaikah. "Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia". In: *Journal of Banking and Finance* 16 (1992), pp. 197–210.

[17] John Y. Campbell, Sanford J. Grossman, and Jiang Wang. "Trading Volume and Serial Correlation in Stock Returns". In: *The Quarterly Journal of Economics* 108.4 (1993), pp. 905–939.

[18] Simon Carter, Wouter Weerkamp, and Manos Tsagkias. "Microblog language identification: overcoming the limitations of short, unedited and idiomatic text". English. In: *Language Resources and Evaluation* 47.1 (2013), pp. 195–215.

[19] M. Cha, H. Haddadi, F. Benevenuto, and P.K. Gummadi. "Measuring User Influence in Twitter: The Million Follower Fallacy." In: *Proc. of ICWSM*. Washington D.C., District of Columbia, 2010.

[20] David Corney, Carlos Martin, and Ayse Gker. "Spot the Ball: Detecting Sports Events on Twitter". In: *Advances in Information Retrieval*. Vol. 8416. Lecture Notes in Computer Science. 2014, pp. 449–454.

[21] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. "#Earthquake: Twitter as a Distributed Sensor System". In: *Transactions in GIS* 17.1 (2013), pp. 124–147.

[22] Aron Culotta. "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages". In: *Proc. of the First Workshop on Social Media Analytics*. Washington D.C., District of Columbia, 2010.

[23] Dmitry Davidov, Oren Tsur, and Ari Rappoport. "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". In: *Proc. of the 23rd International Conference on Computational Linguistics*. Beijing, China, 2010.

[24] Everton Dockery and Manolis G. Kavussanos. "A Multivariate Test for Stock Market Efficiency: The Case of ASE". In: *Applied Financial Economics* 11 (2001), pp. 573–579.

[25] Jingfei Du, Hua Xu, and Xiaoqiu Huang. "Box office prediction based on microblog". In: *Expert Systems with Applications* 41.4, Part 2 (2014), pp. 1680 –1689.

[26] Eugene F. Fama. "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *Journal of Finance* 25.2 (1970), pp. 383–417.

[27] Eugene F. Fama. "Efficient Capital Markets: II". In: *The Journal of Finance* 46.5 (1991), pp. 1575–1617.

[28] Eugene F. Fama and Kenneth R. French. "Permanent and Temporary Components of Stock Prices". In: *The Journal of Political Ecomomy* 96 (1998), pp. 246–273.

[29] A. Ronald Gallant, Peter E. Rossi, and George Tauchen. "Stock Prices and Volume". In: *Review of Financial Studies* 5.2 (1992), pp. 199–242.

[30] Alec Go, Richa Bhayani, and Lei Huang. *Twitter Sentiment Classification using Distant Supervision*. Tech. rep. Stanford University.

[31] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. "Predicting Popular Messages in Twitter". In: *Proc. of International Conference Companion on World Wide Web (WWW)*. Hyderabad, India, 2011.

[32] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. "Social Networks that matter: Twitter under the Microscope". In: *First Monday* 14.1 (2009).

[33] John C Hull. *Options, Futures and Other Derivatives*. 8th. Prentice-Hall, 2011.

[34] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. "Why we Twitter: understanding microblogging usage and communities". In: *Proc. of WebKDD and SNA-KDD Workshop on Web Mining and Social Network Analysis*. San Jose, CA, 2007.

[35] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. "Target-dependent Twitter Sentiment Classification". In: *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Portland, Oregon, 2011.

[36] Charles M. Jones, Gautam Kaul, and Marc L. Lipson. "Transactions, Volume, and Volatility". In: *Review of Financial Studies* 7.4 (1994), pp. 631–651.

[37] Daniel Kahneman. *Thinking, Fast and Slow*. Reprint. Farrar, Straus and Giroux, 2013.

[38] Daniel Kahneman and Amos Tversky. "Prospect theory: An analysis of decision under risk". In: *Econometrica* 47 (1979), pp. 263–291.

[39] N. Kanungsukkasem, P. Netisopakul, and T. Leelanupab. "Recognition of NASDAQ stock symbols in Tweets". In: *Proc. of International Conference on Knowledge and Smart Technology (KST)*. Chonburi, Thailand, 2014.

[40] Jonathan M. Karpoff. "The Relation Between Price Changes and Trading Volume: A Survey". In: *The Journal of Financial and Quantitative Analysis* 22.1 (1987), pp. 109–126.

[41] E. Kouloumpis, T. Wilson, and J. Moore. "Twitter Sentiment Analysis: The Good the Bad and the OMG". In: *Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona, Spain, 2011.

[42] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. "A few chips about Twitter". In: *Proc. of Workshop on Online Social Networks (WOSN)*. Seattle, WA, 2008.

[43] Andrew W. Lo and Jiang Wang. "Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory". In: *The Journal of Finance* 61.6 (2006).

[44] M. Makrehchi, S. Shah, and Wenhui Liao. "Stock Prediction Using Event-Based Sentiment Analysis". In: *Proc. of IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Warsaw, Poland, 2013.

[45] Huina Mao, Scott Counts, and Johan Bollen. "Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data". In: *arXiv:1112.1051* (2011).

[46] Robert McDonald and Daniel Siegel. "The Value of Waiting to Invest". In: *The Quarterly Journal of Economics* 101.4 (1986), pp. 707–727.

[47] L. G. McMillan. *Options as a Strategic Investment*. 5th. Prentice-Hall, 2012.

[48] Robert C Merton. "Optimum consumption and portfolio rules in a continuous-time model". In: *Journal of Economic Theory* 3.4 (1971), pp. 373 –413.

[49] A. Mittal and A. Goel. "Stock Prediction Using Twitter Sentiment Analysis". In: *Proc. of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Washington D.C., District of Columbia, 2013.

[50] M. F. M. Osborne. "Periodic Structure in the Brownian Motion of Stock Prices". In: *Operations Research* 10.3 (1962), pp. 345–379.

[51] Alexander Pak and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In: *Proc. of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta, 2010.

[52] Tushar Rao and Saket Srivastava. "Analyzing Stock Market Movements Using Twitter Sentiment Analysis". In: *Proc. of the 2012 International Conference on Advances in Social Networks Analysis and Mining*. Washington, DC, USA, 2012.

[53] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. "Influence and Passivity in Social Media". In: *Proc. of ACM International World Wide Web Conference (WWW)*. Raleigh, North Carolina, 2010.

[54] Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. "Correlating financial time series with micro-blogging activity". In: *Proc. of ACM International Conference on Web Search and Data Mining (WSDM)*. Seattle, WA, 2012.

[55] Hassan Saif, Yulan He, and Harith Alani. "Semantic Sentiment Analysis of Twitter". In: *Proc. of the 11th International Conference on The Semantic Web - Volume Part I*. ISWC'12. Boston, USA, 2012.

[56] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors". In: *Proc. of International Conference on World Wide Web (WWW)*. Raleigh, NC, 2010.

[57] Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. "Evaluating Sentiment in Financial News Articles". In: *Decision Support Systems*. 53.3 (2012), pp. 458–464.

[58] S. S. Shapiro and M. B. Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3-4 (1965), pp. 591–611.

[59] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. "Exploiting Topic based Twitter Sentiment for Stock Prediction". In: *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, 2013.

[60] Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A. Smith. "Predicting the NFL using Twitter". In: *arXiv:1310.6998*. 2013.

[61] Timm O. Sprenger, Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe. "Tweets and Trades: The Information Content of Stock Microblogs". In: *European Financial Management* 20.5 (2014), pp. 926 –957.

[62] *StockTwits*. http://stocktwits.com.

[63] Tao Sun, Ming Zhang, and Qiaozhu Mei. "Unexpected relevance: An empirical study of serendipity in retweets". In: *Proc. of international AAAI Conference on Weblogs and Social (ICWSM 13)*. Boston, USA, 2013.

[64] Chenhao Tan, Lillian Lee, and Bo Pang. "The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter." In: *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, USA, 2014.

[65] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape". In: *Social Science Computer Review* (2010).

[66] *Twitter Reports Third Quarter 2015 Results. Twitter. 2015-09-30.*

[67] *Twitter Streaming API*. https://dev.twitter.com/streaming/overview.

[68] *Wikipedia*. http://en.wikipedia.org/wiki/GICS/.

[69] Paul Wilmott. *Paul Wilmott on quantitative finance*. John Wiley & Sons, 2013.

[70] *Yahoo! Finance*. http://finance.yahoo.com/.

[71] Charles C. Ying. "Stock Market Prices and Volumes of Sales". In: *Econometrica* 34.3 (1966), pp. 676–685.

[72] L. Zhang. "Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation". In: *Unpublished honor thesis, The University of Texas at Austin*. 2013.

[73] Xue Zhang, Hauke Fuehres, and PeterA. Gloor. "Predicting Asset Value through Twitter Buzz". In: *Advances in Collective Intelligence* 113 (2012), pp. 23–34.

[74] Xue Zhang, Hauke Fuehres, and Peter A. Gloor. "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"". In: *Social and Behavioral Sciences"* 26.0 (2011), pp. 55 –62.