Stony Brook University

Influence Propagation Modeling and Applications in Finance

A Dissertation Presented

by

Li Zhang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Quantitative Finance)

Stony Brook University

May 2016

Stony Brook University

The Graduate School

Li Zhang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Yuefan Deng – Advisor
Professor, Department of Applied Mathematics and Statistics

Svetlozar Rachev – Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Haipeng Xing – Committee Member
Associate Professor, Department of Applied Mathematics and Statistics

Keli Xiao – Outside Committee Member
Assistant Professor, College of Business

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Influence Propagation Modeling and Applications in Finance

by

Li Zhang

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Quantitative Finance)

Stony Brook University

2016

To appropriately use the information from social media for finance-related problems is typically challenging to both finance and data mining. Traditional schemes in finance focus on identifying the trading activities and financial events that generate asset abnormal returns, while the usage of data typically only covers regular events such as earning announcements, financial statements, and new stock issuance. Related data-driven implementations mainly focus on developing trading strategies using social media data, while the results usually lack theoretical explanations. This work is designed to fill the gap between the usage of social media data and financial theories, with comprehensive evaluations using real-world data from social media and the stock market.

A Degree of Social Attention (DSA) framework is developed based on a newly proposed influence propagation model, by leveraging on the vast social networks data, to bring profound impacts on research and practice in finance including market efficiency analysis. For each stock,

the framework dynamically generates a DSA measurement that would accurately reflect the price shock. Specially, the topological structure of a social network is able to be modeled as well as the self-influence of each social media user. Furthermore, the market influence of the current DSA as well as the effects of historical ones on different stocks are estimated.

The essential relationship is verified between social media activities and the stock market movement by testing the semi-strong-form efficient market hypotheses. And then it is confirmed that the effectiveness of our framework in the implementation of stock shock ranking. The results suggest that considering historical DSAs improve the model's performance of fitting abnormal returns in terms of the statistical significance as well as the ranking accuracy. I also develop a new method to estimate social attention of stocks with sentiment analysis and the results show that the newly proposed measurement of social attention would significantly improve the forecasting power of our framework.

This dissertation is dedicated to my parents.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to express my deepest gratitude to my advisor, Professor Yuefan Deng, for these years' guidance, support, encouragement and provision that benefited me much in the completion of this PhD study at Stony Brook University. Thank him so much for guiding me through this Doctoral study process. He also provides us for excellent opportunities to visit scientific conferences and talk to expert in every fields. I am proud to be one of his student for all these time spent with him.

I would like to thank Professor Keli Xiao for his guidance in doing researches of data mining field and all the constructive comments to finish manuscripts. Working with him is the best experience I have in my five years' study as a graduate student. He is a good professor when doing research with us and also a good friend who we can share everything. And I also thank him for helping with my dissertation and being on the committee.

I am very grateful to have Professor Svetlozar Rachev, as the chairperson and Professor Haipeng Xing being in my dissertation committee. I thank them for reading through my research, and providing feedbacks. They give me many expert advices which help a lot with my dissertation. Many thanks to my colleagues and friends for all the scientific discussions and selfless helps. We also have shared lots of happy hours in the work as well as the private times day by day. Finally, I would like to express my love and thanks to my parents and all the family members. It is their endless love and supports which give me all the power to get my PhD degree.

Chapter 1 Introduction

1.1 Motivation

In this research work, I investigate the relationship between social media activities and the stock market movements based on influence modeling and the efficient-market theory. As a popular research topic in cross-fields of finance and data mining, many studies have been conducted for solving related problems, while limitations exist at the same time. In finance, relevant studies based on finance theories, such as asset pricing, market efficiency, and microstructure, are usually limited in terms of the usage of data. On the other hand, although data-driven techniques focus on developing predictive models based on complex social network dataset, the results are usually lack of scientific explanations, and the usage of data is insufficiently supported by theoretical analysis.

In the theory of asset pricing, a price shock can be measured by the abnormal return, which is defined by the difference between the actual return of an asset and the expected one. The innovation of the Internet technologies provides alternative ways to detect these shocks and identifying the underlying causations in addition to traditional approaches. **Figure 1** illustrates the potential connection between the absolute value of abnormal returns and the number of discussions about a Chinese stock (600643) in social media. As highlighted in the figure, the number of discussions matched some jumps in stock abnormal returns but not all of them. In this work, I try to deeply investigate the connection and its implementation value.

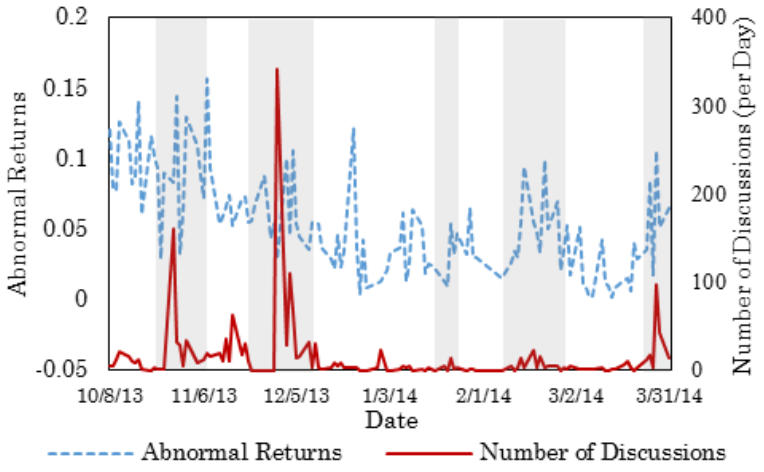


Figure 1. The Abnormal Returns and Number of Discussions for Stock 600643

Since the efficient-market hypothesis was formally introduced by [1], it was frequently studied and became one of the fundamental research topics in finance. In the three forms of market efficiency hypothesis (weak-form, semi-strong-form, and strong-form), the strong-form is rarely studied because of its strict assumption that suggest prices should reflect all public and private information. In the weak-form market efficiency, historical prices do not affect the future; technical analysis does not help in obtaining abnormal returns. Evidence against the weak-form market efficiency can be found in many studies on momentum effect [2-4]. Semi-strong form efficient market hypothesis suggests that market prices should be fully reflected by public information, or abnormal returns will occur. This form of market efficiency is usually investigated based on event studies in order to identify the association between excess returns

and different types of events, such as merger announcement [5], financial reports [6], analyst reports [7] and equity issuances [8]. However, studies on related topics are limited by the traditional event study database in which only standard types of events are included.

The related research works falls into two categories. The first category of studies focuses on identifying the influence of social networks on firm performance and the stock market. The semi-strong-form market efficiency is usually investigated based on event studies in order to identify the association between excess returns and different types of events. Sprenger et al. investigated the relationship between tweet sentiment and stock returns, message volume and trading volume, and disagreement and volatility [9]. Antweiler and Frank discovered a positive relationship has been discovered between disagreement on stock-related articles and the trading volume [10]. Online forums reflect the major activities of uninformed traders, but not informed traders such as institutional investors [11]. Sabherwal et al. claimed that, in terms of the online attention about thinly traded microcap stocks, positive abnormal returns are most likely to be associated with the stocks with the most discussions [12]. Koski et al. showed that noise trading is highly correlated to stock price volatility, while the effect of reverse causation is even stronger [13]. Cha et al. suggested that the actions of retweets and mentions is not solely triggered by followership [14]. Hence, social influence cannot be simply measured by popularity. Boyd et al. claimed that it is a common case that users keep retweeting valuable messages in order to validate nice contents or friend users [15]. Gu et al. believed that there is a tendency that social media users would be attracted by the Internet stock messages with less noise and well processed

contents [16]. Cai et al. found a significant positive relation between stock transaction costs and a company's social ties to the investment community [17]. Faleye et al. investigated social networks of chief executive officer (CEO), and suggested that better-connected CEOs invest more in corporate innovation [18]. Chen et al. studied the extent to which opinions of financial experts shared in social media predict future stock returns and earnings surprises [19].

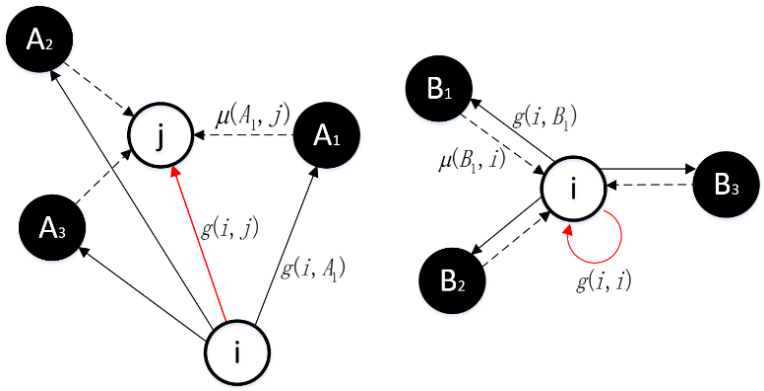
The second category of papers includes the influence models and forecasting models based on social media data. The Independent Cascade (IC) model [20] and the Linear Threshold (LT) model [21] are considered as two of the most famous models in estimating social influence spread [22]. In both models, the influence spread is simply defined as the expected number of activated nodes. With the purpose of stock analysis, however, they may not be able to reflect the true market influence because every node is considered to be equally weighted during the spreading process. To associate social media activities with the stock market, it must model different market influence for social media users. Hayo and Kutan proposed a forecasting model which investigates how other financial markets would affect Russian market [23]. An important variable has been defined in this model: positive or negative news in the past. Although few studies consider the social influence model in stock analysis, social media data has already been used in market stock market prediction. Lavrenko et al. presented a model to predict market behavior based on public event related to target companies [24]. Schumaker and Chen proposed a framework to learn association between news and the stock reactions [25]. Ginsberg, Patel, Brammer, Smolinski, and Brilliant developed a stock prediction model using blog content [26].

In the three forms of efficient-market hypotheses, the weak-form and semi-strong-form market efficiency were studied more frequently. The strong-form is relatively rarely discussed because of its strict assumption that suggest prices should reflect all public and private information. In the weak-form market efficiency, historical prices do not affect the future ones; hence, technical analysis does not help in obtaining abnormal returns. The semi-strong-form efficient market hypothesis suggests that market prices should be fully reflected by the historical prices and other public information. So that fundamental analysis would not generate abnormal returns. This work falls into the investigation of semi-strong-form efficiency, because the finance-related discussions in social media can be categorized as a portion of fundamental analysis. To study the stock market reactions on human activities in social networks, a social influence based framework is proposed that can dynamically capture the social attention for specific stocks in a social network.

On the other hand, an appropriate influence model would result fine estimations of the market influence of user activities in a social network, and lead to effective studies in related topics including price shocks detection. Based on a proposed model [27], the influence of a social media user on a target user is determined by her influence on the neighbors of the target user and the probability of information exchange between the neighbors and the target user.

The model is based on the framework proposed by [27], which introduced a model that considers self-influence of individuals when computing social influence. **Figure 2** shows the influence between two nodes in a social network. Solid lines represent the influence connections, and the

arrows indicate directions of the influence propagation; dash lines with arrows represent the probability of a successful information propagation from one node to another. The main ideas of this work can be summarized as follows. First, the influence of node i on node j is determined by i's influence on j's direct neighbors and their influence propagation to j. Second, the self-influence, which can be considered as the confidence can be less than one (full confidence).



(a) Influence $g(i,j)$ (b) Self-influence $g(i,i)$

Figure 2. Social Networks and Influence Modeling

Although it is believed that the framework would work well in modeling the behaviors of stock market participants, there are still two issues associated with the estimating methods. First, methods for self-influence modeling are insufficiently discussed. Second, the framework does not allow influence to be updated over time. Based on the model, the influence of a social media user on a target user is determined by his/her influence on the neighbors of the target user and the

probability of information exchange between the neighbors and the target user. Moreover, the model is a recursive one that captures the topological structure of a social network. However, there are still some issues associated with this model that can be improved. First, self-influence is an important factor that may significantly affect the influence propagation, but the estimating methods are insufficiently discussed. Second, the framework does not allow influence to be updated over time, while influence propagation should be a dynamic process. To address these problems, a new framework of influence modeling is developed and then apply it to a price shocks detection problem.

This work tries to address the problems and fill the gap between the complex data implementation and finance theoretical analysis in order to link social media activities with the stock market movements. An influence propagation model is introduced with a focus on stock-related articles in social networks, and explore their effects on the stock market based on the efficient-market theory. To have a better understanding about financial market reactions on human activities, I focus on establishing a social influence based framework to, first, by proposing a Degree of Social Attention (DSA) framework, I theoretically analyze the potential relationship between the two systems, and verify its existence by conducting empirical tests on two testable hypotheses derived from the semi-strong-form efficient market hypothesis. Second, the performance of our DSA framework is evaluated in ranking the abnormal returns to verify its ability to capture the price shocks. Both steps would serve as the effectiveness validation of this work. Especially, an influence propagation model based on Gamma distribution for estimating

time effects of previous DSA is proposed. Based on the framework, firstly the relationship between social media and the stock market is confirmed by conducting empirical studies on two hypotheses derived from the efficient market theory. Then, the performance of our DSA framework is evaluated in a ranking problem to verify its ability in capturing stock price shocks.

1.2 Contributions

The main focus of this dissertation is development of Degree of Social Attention (DSA) framework. The main contributions and major results of this work are as follows:

- The improved methods for market influence modeling is composed with the following perspectives. First, most of the existing influence models do not consider the effect of self-influence or confidence during the influence spread estimation. However, in stock analysis, it is believed that the market influence of individuals in a social network is determined by their social relations as well as their confidence in the articles they write, comment, and repost. If we assume people's confidence in specific topics mainly depends on the knowledge and expertise they have in handling related information, then different values must be assigned to describe the differences among people. Second, both the stock market and social media are dynamic systems. To explore the stock market dynamics by social media data, an influence updating procedures is designed to capture the real-time changes in a social media system. Based on the updated influence, the DSA is defined for

each stock. Hence, the two dynamic systems are able to connect together. Last, the computational complexity of influence modeling and challenge of computation is created with the purpose of dynamic stock analysis. To this end, it provides several alternative approaches as well as algorithm designs for computational efficiency.

- Based on the proposed algorithm, the Degree of Social Attention (DSA) is defined for each stock and the effectiveness of DSA as the key factor to link social media activities to the stock market is tested and confirmed. Regarding the relationship between social media and the stock market, it also depends on whether the social media users are informed traders or uninformed traders. It is a common agreement that the Chinese market is empirically inefficient in the Semi-Strong form [12]. This indicates that public information is not fully reflected by the market price. During the trading period, uninformed traders usually cannot process information efficiently, so they simply follow the market, and cause over trading and an increasing amount of abnormal returns. On the other hand, the trading activities of informed traders would pull the market price back to the intrinsic value, and reduce the abnormal return. Thus, two testable hypotheses are proposed and based on the test results from selected samples of highly active stocks in the Chinese market, it is verified the significant effect of DSA on stock abnormal returns, and would further serve as a new evidence to support the semi-strong-form market inefficiency hypothesis.

- In addition to the DSA at current time, the effect of historical DSA should be considered on the stock market movement. An influence propagation model is proposed based on Gamma distribution to estimate the time effect on historical DSAs. The model, which is used to capture the time effect of historical DSA, improves the performance of our DSA framework in real-world data fitting. The sample is expanded, and more comprehensive validation are conducted by testing the full sample as well as several subgroups separated by business sectors and listing markets. 34 leading stocks in the Chinese market are selected in terms of the average trading volume. The tests are based on minute level high-frequency data. The results are consistent with the previous one, and also statistically confirmed the effectiveness of the newly proposed DSA framework. Instead of separately testing each stock, new tests are based on the aggregate sample, and robustness checks are conducted based on subsamples of business sectors and listing markets; hence the new results are considered as more statistically reliable ones. To evaluate the effectiveness of the newly proposed DSA framework in financial implementation, a ranking problem is also applied for price shocks detection. By comparing with several baseline schemes, it is confirmed that the new method significantly improves ranking performances.

- Many recent studies on finance and social media discovered that investor's attention is significantly correlated to the financial market movement in terms of the price shocks. Following related findings, a significant and challenging problem is to forecast the direction of the market movement based on vast social media activities. Appropriately processing social networks data and developing models to capture investors' attention on stocks may effectively help in forecasting tasks. I propose and then apply a price shocks forecasting framework, which simultaneously takes social influence of social network users and their opinions into consideration. Specifically, a new method is developed to estimate social attention of stocks with sentiment analysis. Moreover, the effect of historical market information on the future movement is considered. Based on a series of tests on the Chinese stock market, the effectiveness of our framework is verified. The results also show that the newly proposed measurement of social attention would significantly improve the forecasting power of our framework.

Chapter 2 Influence Modeling and Degree of Social Attention (DSA)

The goal of this work is to detect the stock market movement based on social media activities. To do, first thing is to identify the essential connection between the two systems. That requires to carefully select and extract variables from both side. For the stock market, its movement can be represented by the periodic returns and abnormal returns. Then, the major challenge of the problem is to generate a social media activity measurement that can appropriately describe human's reactions to finance-related information. The problem can be formalized as follows.

Given a dynamic social network $\mathcal{G} = (\mathcal{N}, E)$, where $\mathcal{N}$ is a set of nodes (social media users) and E is a set of edges, $\mathcal{I}$ the set of historical information flows within the network, and the return $\{R_{q,t}\}$ for stock $q = 1, 2, 3, \dots Q$ and time $t = 1,2, \dots, T$, the goal is to generate a social attention measurement with a purpose of detecting the price movement of a targeting stock or portfolio. The measurement needs to satisfy several requirements as follows. First, to improve the social attention measurement proposed in [28], an earlier phase of this work, the new measurement should simultaneously take the current social attention and the historical ones into consideration. Second, the framework should be able to capture the time effect of influence propagation. Finally, the effectiveness of our approach should be confirmed by 1) significance tests designed under the theory of market efficiency and 2) the experimental results from the implementation of financial shocks detection.

In this chapter, I will first talk about the influence modeling by showing the typical influence model and the updating process with self-confidence modeling. Next I will introduce the Degree of Social Attention (DSA) framework for stock analysis. Lastly I will talk about the efficient approaches which several alternative ways are provided and discussed for computing performance and modeling purpose.

2.1 Influence Modeling

2.1.1 Social Influence Modeling

The example of social networks and the influence modeling process in this work can be seen in **Figure 2**. Solid lines represent the influence connections, and the arrows indicate directions of the influence propagation; dash lines with arrows represent the probability of a successful information propagation from one node to another. For instance, $\mu(A_1, j)$ is the probability of influence is propagated from node A_1 to j. Let $\mathbf{A} = \{A_1, A_2, \dots A_n\}$ be the set of neighbors of node j, i's influence on j, $g(i,j)$, can be measured by the sum of the weighted-influence of i with the assessment of **A** [27]. To solve this recursive function, an initial value should be assign to $g(i,i)$, the self-influence of i.

Based on influence model developed by [27], the influence of node i on j, it can be mathematically defined as:

$$g(i,j) = \begin{cases} \frac{1}{1+\lambda_j} \sum_{k \epsilon \mathcal{N}_j} g(i,k)\mu(k,j), & \text{if } i \neq j \\ e_i, & \text{if } i = j \end{cases} \quad (1)$$

where $\mathcal{N}_j$ is $j's$ trust-friend set. If $k \epsilon \mathcal{N}_j$ then j and k are connected. And $\mu(k,j)$ is the propagation probability form k to j. It can be measured by the probability that j will take actions on an article posted by k. For a special case, the propagation probability from i to itself, $\mu(i,i)$ is set to 1. Parameter λ_j is the discount factor of j that measures the influence diminishing during propagation. I follow [27], and set an identical λ for all nodes. $e_i \in [0,1]$ is the prior constraint value being assigned to each node i. If i is fully confident, this value is assigned to be one; if i has no confidence at all, it would be zero.

Then, the social influence of i is defined as follows:

$$G(i) = \sum_{j=1}^{N} g(i,j) \quad (2)$$

where N represents the total number of users in the social network.

To solve the problem, rewrite **Equation (1)** as:

$$g(i,j) = \frac{1}{1+\lambda} \sum_{k \epsilon \mathcal{N}} g(i,k)\mu(k,j) + v_{ij} \quad (3)$$

Notice that $\mathcal{N}$ equals to $\mathcal{N}_j$ because $\mu(k,j) = 0$ if $k \notin \mathcal{N}_j$. v_{ij} is the j-th entry in a vector $\mathbf{v_i} = [0,0,\dots,v_{ii},\dots,0]'$, in which the i-th entry v_{ii} ensures $g(i,i) = e_i$. Let $\mathbf{g_i} = [g(i,1), g(i,2), \dots, g(i,n)]'$, based on **Equation (3)**, it can be rewritten in the matrix format:

$$\begin{aligned} \mathbf{g_i} &= (\mathbf{I} + \lambda\mathbf{I})^{-1}(\mathbf{M}'\mathbf{g_i} + \mathbf{v_i}) \\ &= (\mathbf{I} + \lambda\mathbf{I} - \mathbf{M}')^{-1}\mathbf{v_i} \\ &= \mathbf{P} \cdot \boldsymbol{v_i} = \mathbf{P}_{\boldsymbol{i}\cdot}\boldsymbol{v_{ii}} \end{aligned} \tag{4}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{M}$ is a N by N influence transition matrix of $\mu(k,j)$; $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{M}')^{-1}$. In **Equation (4)**, $(\mathbf{I} + \lambda\mathbf{I} - \mathbf{M}')$ is invertible because it is strictly diagonally dominant as $\mu(i,i) = 1$. As $g(i,i) = e_i = v_{ii}p_{ii}$ and v_i only has the i-th entry v_{ii} nonzero, we can get:

$$g(i,j) = \frac{e_i}{p_{ii}}p_{ji} \tag{5}$$

Therefore, **Equation (2)** can be rewritten as:

$$G(i) = \sum_{j=1}^{N} g(i,j) = \frac{e_i}{p_{ii}}\sum_{j=1}^{N} p_{ji} \tag{6}$$

To solve the recursive function $g(i,j)$, an initial value should be assign to e_i, the self-influence of i. Finally, the influence $g(i,j)$ and i's social influence $G(i)$ are computed based on **Equation**

(5) and **Equation (6)** respectively. One important issue that has not been sufficiently discussed in previous work is the definition of the self-influence e_i.

2.1.2 Modeling Self-Confidence

Based on **Equation (1)**, the self-influence e_i has to be given before the influence $g(i,j)$ can be computed. [27] claimed that e_i can be viewed as the confidence level of i, but no method has been provided for the estimation. In this work, we assume that the confidence levels are determined during the interactions among social media users. The confidence increases when a user receives additional feedbacks from the friends. Here we do not consider whether these feedbacks are positive or negative, as either way would lead to an increasing social influence.

As shown in **Figure 2(b)**, the self-influence (confidence) of node i, which can be denoted by $g(i,i)$, depends on 1) how much influence i has on his neighbors, and 2) the probability of the neighbors would react on i's actions. The value of someone's confidence is high when his neighbors give more reactions to the articles posted by him or he has large influence on them. Thus, the confidence of node i is defined as follows:

Definition 1. The confidence of user ***i*** in a social network is the sum of the product of the user's influence on its direct neighbors and the probability of feedback it receives from those neighbors.

It can be mathematically expressed as:

$$g(i,i) = \frac{1}{1+\lambda} \sum_{k \epsilon \mathcal{N}_i} g(i,k)\mu(k,i) \tag{7}$$

where $\mathcal{N}_i$ is i's trust-friend set, $\mu(k,i)$ is the propagation probability from k to i, and $g(i,k)$ is the influence of i on k. Then **Equation (1)** is modified as follows:

$$g(i,j) = \frac{1}{1+\lambda_j} \sum_{k \epsilon \mathcal{N}_j} g(i,k)\mu(k,j) \text{ , for } \forall i,j \in \mathcal{N} \tag{8}$$

This recursive equation can be solved with an initial value of $g(i,i)$ in a dynamic system.

2.1.3 Influence Updating

To capture the dynamic information in a social network, the following influence updating process is applied. As shown in **Algorithm 1**, the confidence for node $i \in \mathcal{N}$ at time t is denoted by:

$$g(i,i;t)=\begin{cases}\dfrac{1}{1+\lambda}\displaystyle\sum_{k\epsilon\mathcal{N}} g(i,k;t)\mu(k,i;t), & \text{for } t>0\\ 1, & \text{for } t=0\end{cases} \tag{9}$$

ALGORITHM 1. Influence Updating

Input: $\lambda, g(i,i;0), t_{now}$

Output: $G(i;t)$ and $\mathbf{g}_{\mathrm{i}}(\mathrm{t})=[g(i,1;t), g(i,2;t), \dots, g(i,N;t)]'$, for $t=1,2,\dots,t_{now}$

$g(i,i;0)=1$;

for ($\boldsymbol{t}=\mathbf{0}$; $\boldsymbol{t}<\boldsymbol{t_{now}}$; $\boldsymbol{t}$++) **do**

  if ($\boldsymbol{t}>\mathbf{0}$)

    $\mathbf{M}_{t-1}=\mathbf{M}_t$; // save the probability matrix for t-1

    $G(i;t-1)=G(i;t)$; // save the influence matrix for t-1

    Update e_i *based on* $\mathbf{M}_{t-1}$ *and* $\mathbf{g}_{\mathrm{i}}(t)$*;*

  end if

  Compute $\mathbf{M}_{\boldsymbol{t}}=[\mu(i,j)]_{n\times n}$ *based on the data of one month prior;*

  for ($\boldsymbol{i}=\mathbf{0}$; $\boldsymbol{i}<\boldsymbol{n}$; $\boldsymbol{i}$++) **do**

    for ($\boldsymbol{j}=\mathbf{0}$; $\boldsymbol{j}<\boldsymbol{n}$; $\boldsymbol{j}$++) **do**

      $g(i,j)=\frac{g(i,i;t)}{p_{ii}}p_{ij}$;

      $\mathbf{g}_{\mathrm{i}}(t)$.pushback($g(i,j)$);

    end for

  end for

  $G(i;t)=\mathrm{sum}(\mathbf{g}_{\mathrm{i}}(t))$;

  return $G(i;t)$, $\mathbf{g}_{\mathrm{i}}(t)$;

end for

$g(i,i;0) = 1$ is set for each node as the initial value by assuming people are initially fully confident. The influence of i on j, $g(i,j)$, and i's total social influence, $G(i)$, are updated over time based on **Equation (5)** and **Equation (6)**. The whole social influence system is updated with e_i based on the information in the previous time window.

Figure 3 shows an example of the estimated self-confidence for a social media user in our sample. As can be seen, the confidence is initialized to one and then decreased over time. Based on *Definition 1*, the fluctuations are determined by the user's current social influence as well as the reactions from her direct neighbors.

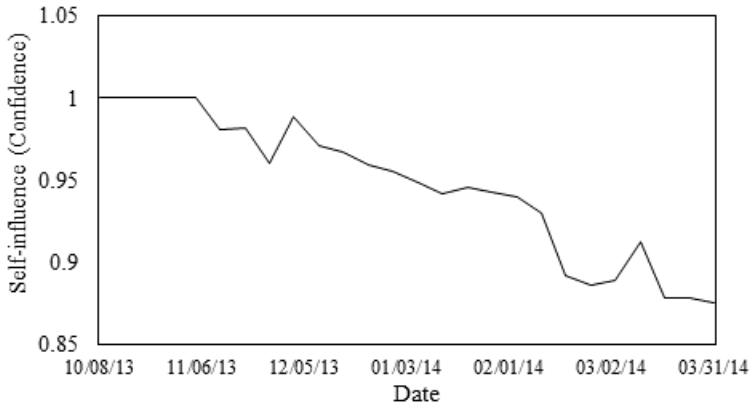


Figure 3. An Example of Dynamic Confidence

An important issue of the algorithm is the determination of the length of time window. Large time windows may result the existence of accumulative outdated information or noise data; small

ones may result biased samples because important historical information can be easily discarded. In this work, the influence is updated weekly based on a time window which covers the past four weeks' social media data. The detailed discussion is provided in chapter 3.

2.2 Degree of Social Attention (DSA)

In this section, the Degree of Social Attention (DSA) framework is introduced for stock analysis. I define the DSA and theoretically analyze its impact on the stock market based on the efficient-market theory.

Social attention, sometimes called investors' attention in finance, refers to the level of notice taken of some specific stocks by people. Based on [28], given a social network $\mathcal{G} = (\mathcal{N}, E)$, for any user $i, j \in \mathcal{N}$, let $g(i, j; t)$ be the influence of i on j, the Degree of Social Attention (DSA) to stock q at time t is defined as:

$$a_q(t) = \sum_{i=1}^{N} \sum_{j=1}^{N} \varphi(i, j; t) d_q(i; t) f_q(i, j; t) \tag{10}$$

where N represents the total number of users in the social network; $d_q(i; t)$ is the number of articles posted by i at time t; $d_q(i; t) = 0$ if i generate relevant information about stock q at time t. $\varphi(i, j; t)$ is a discount factor that is related to the characteristics of i and j at time t. Possible characteristics include age, education level, title, gender, location number of followers,

frequency of article posting, and more. In this work, because of the limitation to access related information, we assume $\varphi(i,j;t)$ to be one for all i and j.

Then, the current DSA to stock q within a time interval $[t_s, t_e]$ is defined as follows:

$$\mathcal{A}_q(t_s \to t_e) = \int_{t_s}^{t_e} a_q(t)\, dt \tag{11}$$

2.3 Efficient Approaches

One challenge of implementing the influence updating process is the problem of high computational complexity. The algorithm contains a process of N by N matrix inversion, $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{\Gamma}')^{-\mathbf{1}}$ for each time t. (The propagation probability $\mu(i,j)$ which forms the propagation matrix Γ can be calculated as the probability that node j reacts on node i's posts - comment, repost, like - during each time period) The complexity of matrix inversion with preferred methods, such as Gaussian Elimination, is $O(N^3)$. While it is difficult to reduce the computational complexity of inverting matrix, two approaches are provided to reduce matrix dimension with the purpose of stock analysis.

2.3.1 Market-based Approach

Although there are a huge number of users in the whole social network, it is not necessary to include all of them for stock analysis. It believes that only those who participate in the discussions of stock-related topics can influence the market. Furthermore, it can be true that people are only interested in stocks listed on the same market, say, NYSE or Nasdaq. In this case, if we separately consider the users who discuss stocks from different markets, we would only need to handle a matrix with smaller size for each market. Therefore, the influence modeling is modified as follows:

$$g_\theta(i,j) = \frac{1}{1+\lambda} \sum_{k \epsilon \mathcal{N}_\theta} g_\theta(i,k)\mu_\theta(k,j) \text{ , for } \forall i,j \in \mathcal{N}_\theta \qquad (12)$$

$$\theta = selected\ market$$

where the size of θ is the number of stock markets.

While this approach significantly reduces the size of matrix **P** by separately considering each market, the limitations are still obvious. First, it only works when the overlapping among these markets is small. If we find that most of people are interested in all markets, the size would not be reduced much. Second, there are usually no more than three stock markets within a country. For instance, there are only two major stock markets in the U.S. (NYSE and Nasdaq). Therefore, the optimal case is to reduce the matrix size to 1/3 with which the complexity is still too high.

2.3.2 Stock-based Approach

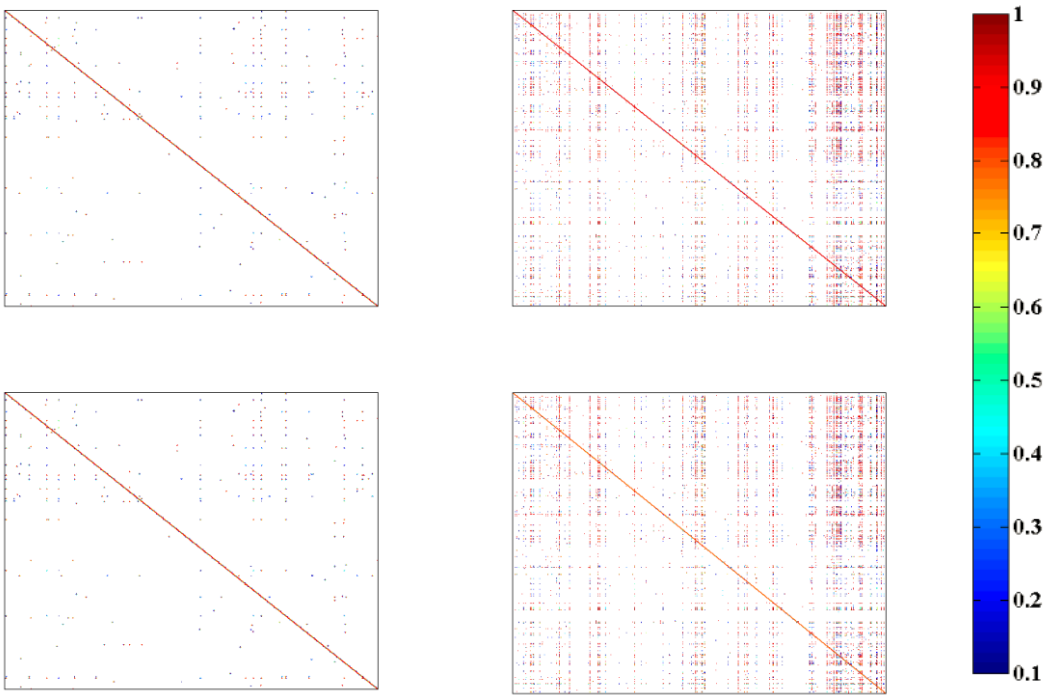
To further reduce the computational cost, a stock-based approach is illustrated which considers each single stock as an independent system. Similar to the market-based approach, we assume people are interested in a small group of particular stocks in a certain time but focusing on all of them. So that the modeling process can be separate implemented based on data of each stock. We follow the same process of **Equation (12)** where $\boldsymbol{q}$ links to a selected stock but market as follow:

$$g_q(i,j) = \frac{1}{1+\lambda} \sum_{k \epsilon \mathcal{N}_q} g_q(i,k)\mu_q(k,j) \text{ , for } \forall i,j \in \mathcal{N}_q \tag{13}$$

$$q = selected\ stock$$

where the size of q is the number of stocks.

For the influence computation, $\lambda = 0.176$ is set as suggested by [27, 28]. The influence matrices of SSE market and stock 600688 are plotted in **Figure 4** as two examples showing how influence matrix $\mathbf{G} = [g_q(i,j)]_{n \times n}$ is updated over time. As can be seen, **F** is an asymmetry matrix which indicates that $g_q(i,j) \neq g_q(j,i)$. The diagonal values are the estimated self-influence or confidence. Blank areas indicate zero-influence between users. **Figure 4(a)** and **Figure 4(b)** plots the influence matrices of Shanghai Stock Exchange (SSE) market and stocks 600688.



(a) Influence Matrix (Market-based)

(b) Influence Matrix (Stock 600688)

Figure 4. Potential Diagram of the Influence Matrix for (a) SSE Market and (b) Stock 600688, at 11/4/2013 (upper) and 1/13/2014 (lower)

The influence matrix for the market-based sample **Figure 4(a)** has more blanks than the one generated by stock-based approach **Figure 4(b)**. Hence, it can conclude that most users in a social network only focus on several specific stocks instead of the whole market, and the influence matrices generated by the stock-based approach should be effective for stock analysis.

2.3.3 Algorithm Parallelization

The market-based and stock-based approach are designed to reduce the size of matrix **P**, hence reduces the computational cost. The generalized minimal residual method (GMRES) [29] with QR factorization method is used to solve this matrix inversion problem $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{M}')^{-\mathbf{1}}$, as shown in **Algorithm 2**. Moreover, since the divided tasks from the two approaches are independent with each other, parallel computing techniques can be easily applied to enhance the computational efficiency.

Now I am going to talk about the study of computational efficiency of DSA framework. The major computational cost of the framework is the dynamic influence modeling process, so it needs to compare the performance of the market-based approach and the stock-based approach with a benchmark algorithm that considered all social media users as a single group. The experiments are based on synthetic data for larger sample size. According to the statistics of the 6-month social media data that covers 100% user information of the stocks in sample, the average number of users for one stock is 2,177.31. 2,000 social media users are assigned for each stock to run the tests. The propagation probabilities for each pair of users are randomly determined for sample tests. Currently there are 2,314 stocks trading in the Chinese stock market, and the maximum stock number is set as 2,000 for the experiments.

ALGORITHM 2. GMRES(K) with QR Factorization Method

Input: $\lambda, k, \mathbf{M}$

Output: $\mathbf{P} = (\mathbf{I} + \lambda\mathbf{I} - \mathbf{M}')^{-1}$

$\mathbf{A} = \mathbf{I} + \lambda\mathbf{I} - \mathbf{M}'$ // $\mathbf{A} * \mathbf{P} = \mathbf{I}$

// QR factorization

$\mathbf{A} = \mathbf{Q} * \mathbf{R}$ // $\mathbf{R}$ is a upper triangular matrix; $\mathbf{Q} * \mathbf{Q}' = \mathbf{I}$; $\mathbf{R} * \mathbf{P} = \mathbf{Q}'$

for ($\boldsymbol{i} = \mathbf{1}$; $\boldsymbol{i} \leq \boldsymbol{N}$; $\boldsymbol{i}$++) // GMRES solve $\mathbf{P}_i$ for $\mathbf{R} * \mathbf{P}_i = \mathbf{Q}_i'$

$\mathbf{P}_i = \mathbf{0}$

$\boldsymbol{r} = \mathbf{Q}_i' - \mathbf{R} * \mathbf{P}_i$

do while $\| \boldsymbol{r} \| > \mathbf{1}\boldsymbol{e}^{-6}$

$\boldsymbol{v_1} = \boldsymbol{r} / \| \boldsymbol{r} \|$

for ($\boldsymbol{j} = \mathbf{1}$; $\boldsymbol{j} \leq \boldsymbol{k}$; $\boldsymbol{j}$++)

for ($\boldsymbol{s} = \mathbf{1}$; $\boldsymbol{s} \leq \boldsymbol{j}$; $\boldsymbol{s}$++)

$\boldsymbol{h_{s,j}} = (\mathbf{R}\boldsymbol{v_j})'\boldsymbol{v_s}$

end for

$\tilde{\boldsymbol{v}}_{\boldsymbol{j+1}} = \mathbf{R}\boldsymbol{v_j} - \sum_{\boldsymbol{s=1}}^{\boldsymbol{j}} \boldsymbol{h_{i,j}}\boldsymbol{v_i}$

$\boldsymbol{h_{j+1,j}} = \| \tilde{\boldsymbol{v}}_{\boldsymbol{j+1}} \|$

$\boldsymbol{v_{j+1}} = \tilde{\boldsymbol{v}}_{\boldsymbol{j+1}} / \boldsymbol{h_{j+1,j}}$

end for

solve: $\min_{\boldsymbol{y}} \left\| \|\boldsymbol{r}\|\boldsymbol{e_1} - \boldsymbol{H_j}\boldsymbol{y_j} \right\|$ for $\boldsymbol{y_j}$ // where $\boldsymbol{H_j} = [\boldsymbol{h_{1,j}}, \boldsymbol{h_{2,j}} \ldots \boldsymbol{h_{j,j}}]$, $\boldsymbol{e_1} = [\mathbf{1}, \mathbf{0} \ldots \mathbf{0}]'$

$\mathbf{P}_i = \mathbf{P}_i + \boldsymbol{V_j}\boldsymbol{y_j}$ // where $\boldsymbol{V_j} = [\boldsymbol{v_1}, \boldsymbol{v_2} \ldots \boldsymbol{v_j}]$

$\boldsymbol{r} = \mathbf{Q}_i' - \mathbf{R} * \mathbf{P}_i$

end do

end for

return $\mathbf{P}$

Figure 5 plots the average computational time for the three approaches based on 10 fair experiments. As it can be seen, the stock-based approach performs the best among the three approaches. A slight improvement is also found for the market-based approach. Furthermore, the stock-based approach creates opportunities of parallel computing. **Figure 6** plots the speedup for different matrix dimension with parallel computing. With more cores, larger speedup and better computational efficiency can be obtained. For instance, for 4 million nodes, it can get 5.52 times speedup by using 8 cores, and the speedup achieve 8.14 times by using 16 cores. For this algorithm, the speedup is very efficient due to small communication cost among cores. One issue is the reduction of speedup performance when increasing the number of nodes. For example, when to handle 400k nodes with 16 cores, it gives 10.29 times speedup, while handling 4 million nodes reduces the speedup to 8.14 times. However, the total time consuming in the analysis is controlled in a tolerable level with good scalability using high-performance computing methods.

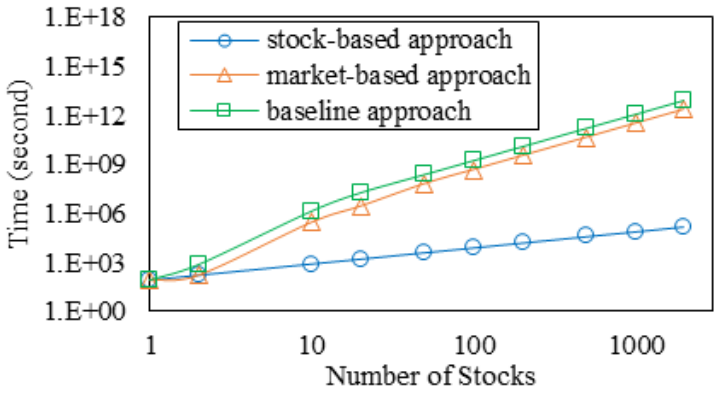


Figure 5. Computational Costs of Three Approaches: Solving One Influence Matrix with Different Stock Number

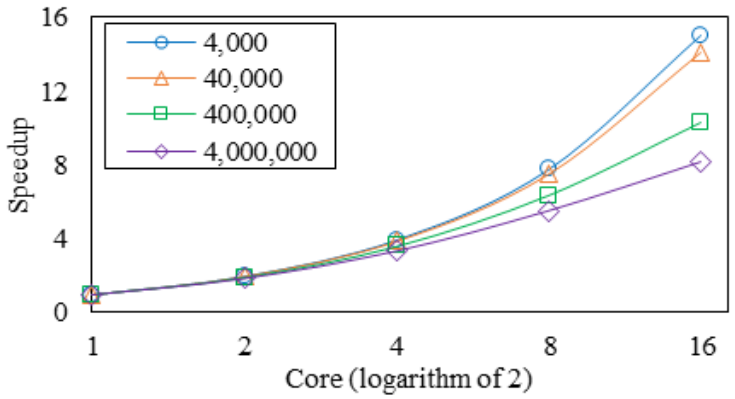


Figure 6. Speedup Results for Four Matrix Dimension (4k, 40k, 400k, 4m Nodes) with Different Computing Cores

Chapter 3 Statistical Hypothesis Tests

The methodology of influence modeling and alternative approaches is discussed in the previous chapter. In this chapter, the model is evaluated based on one major consideration, to investigate whether the Degree of Social Attention (DSA), as the major index proposed in this work, has significant association with abnormal returns [28].

First, the computation of stock abnormal return is introduced and two testable hypotheses are proposed for significance tests accordingly. Then it shows the data structure in this research work and followed by the sample analysis of 10 specific stocks with time windows determination. Several regression models are designed based on the hypotheses and results analysis is shown in the last section.

3.1 Stock Abnormal Return

As a typical setting in finance, stock traders are separated into two types: informed traders and uninformed traders [30]. Informed traders have significant advantages during the trading in terms of specialized information, technical skills, and capital power. Uninformed traders are considered as noise traders who trade on what they think is information but in fact is merely noise. Theoretically, when assets are mispriced, the activities of informed traders would pull prices back to fundamental values, and abnormal returns are reduced. The activities of uninformed traders would generate noise, and increase stock abnormal returns, regardless whether the

information is positive or negative. Therefore, to explore the relationship between social media and the stock market, it need to be figured out whether the social media users are informed traders or uninformed traders. And stock abnormal returns must be the key to identify such relationship.

In finance, an abnormal return, or price shock, can be defined as the difference between the actual return and the expected one [31]. For a stock q, its abnormal return at time t is:

$$AR_{q,t} = R_{q,t} - E_t[R_{q,t}] \tag{14}$$

where $AR_{q,t}$ is the abnormal return of stock q at time t; $R_{q,t}$ is the actual return of the stock; and $E_t[R_{q,t}]$ is the expected return. To estimate the expected return, firstly the beta risk (systematic risk), β_q, is estimated based on an ordinary least squares (OLS) regression [32, 33] as follows:

$$R_{q,t} = \theta + \beta_q R_{m,t} + \varepsilon_{q,t} \tag{15}$$

where θ is a constant term; $R_{m,t}$ is the market return which can be computed based on market index. The systematic risk, β_q, is estimated as the coefficient of the market return. $\varepsilon_{q,t}$ is the error term. Assume the risk-free rate at time t is $R_{f,t}$, the expected return for stock q can be computed based on the famous Capital Asset Pricing Model (CAPM) [34, 35]:

$$E_t[R_{q,t}] = R_{f,t} + \hat{\beta}_q\left(R_{m,t} - R_{f,t}\right). \tag{16}$$

The initial question to answer is whether DSA can be directly used in stock forecasting. Thus, the first testable hypothesis is proposed as follows:

Hypothesis 1. There is no significant relationship between the Degree of Social Attention (DSA) and stock returns if uninformed traders dominate the social media activities; the relationship may exist if informed traders are in dominant positions.

It is a common agreement that the Chinese market is empirically inefficient under the semi-strong form [36]. It indicates that public information is not fully reflected by the market price. Generally, uninformed traders cannot effectively process publicly available information as informed traders do. Their trading behaviors will result an increasing of abnormal return. On the other hand, the trading activities of informed traders reduce the abnormal return. Thus, our second hypothesis is formulized as:

Hypothesis 2. Uninformed traders dominate social media activities, and hence there is a positive relationship between the Degree of Social Attention (DSA) and the absolute value of abnormal returns.

Both hypotheses need be tested statistically, and then it can conclude the essential connection between social media and the stock market.

3.2 Data Processing

Stock-related data is collected from two major sources: the public social media data and the stock market data. This work focuses on the Chinese stock market - Shenzhen Stock Exchange (SZSE) and Shanghai Stock Exchange (SSE) - and the social media activities in Weibo.com, the largest mobile social network in China.

3.2.1 Social Media Data

To collect data from Weibo.com, it program based on an open source tool HtmlUnit [37], which can model HTML documents of Weibo, then identify and retrieve information needed.

Table 1 summarizes the collected features of social media data. Three types of features are collected for stock analysis. The first one contains the basic attributes of each posted article including the article ID, author account ID, the content, and the date and time for posting. The second one contains the features measuring social reactions, such as the number of times of the article being ‘like’, ‘repost’, and ‘comment’. The last type of features includes the IDs of reactions, which is used to track the characteristics of each participator.

Table 1. Weibo Data Profiling

Data	Feature Descriptions
Basic Identifications	Weibo ID Account ID Post Content Date and Time
Influence-Related	Number of "like" Number of Reposts Number of Comments
Reaction Tracking	Reaction ID

Table 2 reports the key statistics of our experimental data. The full sample contains six-month data from October 2013 to March 2014. It includes stock-related articles of 34 selected stocks listed on SZSE and SSE. Influence modeling is implemented based on the data of the users who have stock-related activities. The information of other users is discarded, because we assume they are not stock participants even if some of them have large social influence, their influence on the stock market is considered to be zero.

One question raised is how to identify informed and uninformed traders. Based on related literature, informed traders are usually more active in terms of trading frequency [38, 39] and have more market influence than unformed traders [30]. **Figure 7** shows the distribution of the

sample social influence $G(i)$. Among the accounts associated with selected highly active stocks in our sample, the distribution of the average social influence of i is computed. As can be seen, 86.4% social media users are with social influence less than one ($\ln(G(i)) < 1$). They are considered as small social influence participants. The rest 13.6% who have log social influence greater than one are considered as big and medium influence participants. Thus, we assume most of the social media users are uninformed traders because of their small social influence.

Table 2. A Summary of Data Statistics

Data Sources	Properties	Statistics
Common	Time Scale	10/08/2013 – 03/31/2014
	Number of Days	174
Social Media	Number of Posts	139,855
	Number of Accounts	20,410
	Number of Posts per Day	803.76
Stock Market	Number of Stocks	34
	Number of Business Sectors	6
	Number of Trading days	119
	Data Frequency (Hz)	1/600
	Number of Time Points	3,094 per stock

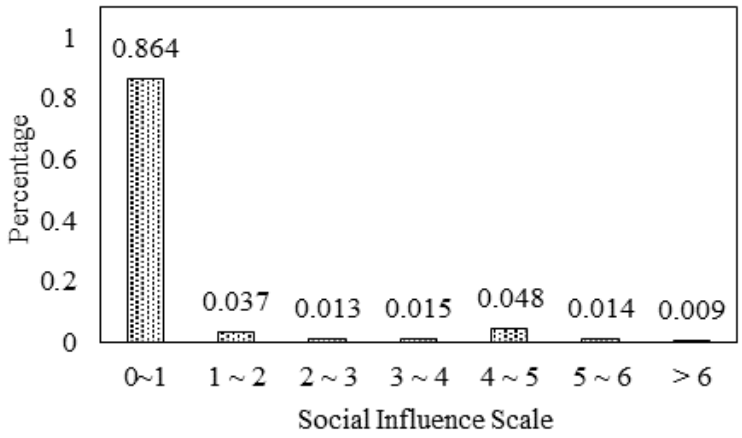


Figure 7. Distribution of the Social Influence $G(i)$

3.2.2 Stock Market Data

Our stock data includes 10-minute stock prices and volumes of 34 highly active Chinese stocks from October 8th 2013 to March 31st 2014. The Shanghai Shenzhen CSI 300 Index is also collected as the market index [40]. All prices are adjusted for dividends and splits. **Table 3** presents the full list of selected stocks in our sample with their business sectors. **Table 4** reports the statistics of trading volumes of the total market as well as the selected 34 leading companies. Our sample is categorized by two listing markets, SSE and SZSE; and six business sectors, including Basic Materials (BM), Consumer Goods (CG), Financial (F), Industrial Goods (IG), Services (S) and Utilities (U). As can be seen, the average trading volumes of the stocks in our sample are significantly larger than the mean trading volume of total market.

Table 3. List of Selected Stocks

(a) SZSE Market						
Stock Code	000002	000031	000100	000157	000554	000598
Company Name	China Vanke	COFCO Property	TCL	Zoomlion Heavy Industry Sci. and Tech.	Sinopec Shandong Taishan Petroleum	Xingrong Investment
Business Sectors	F	F	CG	IG	BM	F
Stock Code	000629	000709	000725	000767	000783	000875
Company Name	Pangang Group Vanadium Titanium & Resources	Hebei Steel	BOE Technology Group	Shanxi Zhangze Electric Power	Changjiang Securities	Jilin Power Share
Business Sectors	BM	BM	CG	U	F	F
Stock Code	002024	002183	002277	002490	300185	
Company Name	Suning Commerce Group	Eternal Asia Supply Chain Management	Hunan Friendship & Apollo Commercial	Shandong Molong Petroleum Machinery	Tongyu Heavy Industy	
Business Sectors	F	F	F	IG	IG	

Table 3. List of Selected Stocks *(Cont.)*

(b) SSE Market						
Stock Code	600048	600060	600157	600169	600221	600317
Company Name	Baoli Real Estate	Hisense Electric	WTECL	Taiyuan Heavy Industry	Hainan Airlines	Yingkou Port Liability
Business Sectors	F	CG	BM	IG	S	S
Stock Code	600643	600688	600795	600863	601018	601106
Company Name	Shanghai Aijian Group	Sinopec Shanghai Petrochemical	GD Power Development	Inner Mongolia Mengdian Huaneng Thermal Power	Ningbo Port Liability	China First Heavy Industries
Business Sectors	F	BM	U	U	S	IG
Stock Code	601118	601608	601618	601899	601989	
Company Name	China Hainan Rubber Industry Group	CITIC Heavy Industries	Metallurgical Corporation of China	Zijin Mining Industry	China Shipbuilding Industry	
Business Sectors	BM	IG	BM	BM	S	

Table 4. Trading Volumes of Selected Samples (Million Shares)

Total Market				
	Max	Mean	Min	Std.
Total Market	0.128	0.008	0.001	0.114
Selected Samples				
	Max	Mean	Min	Std.
SSE	0.116	0.049	0.012	0.059
SZSE	0.113	0.054	0.010	0.060
BM	0.116	0.048	0.016	0.071
CG	0.107	0.047	0.018	0.067
F	0.108	0.049	0.012	0.065
IG	0.102	0.042	0.010	0.051
S	0.111	0.051	0.018	0.067
U	0.115	0.050	0.013	0.083

The Chinese stock market is open every Monday to Friday with two separated sessions: 1) the morning session begins from 9:30 to 11:30; 2) the afternoon session starts from 13:00 to 15:00. To avoid the noise information overnight and during the lunch break from 11:30 to 13:00, the first 10 minutes of each session is removed from our sample. Based on the sample, stock returns are computed as the current change of price divided by the previous price. Abnormal returns are computed based on the approach introduced in section 3.1.

Based on the efficient market theory, sufficient trading activities would reduce the abnormal return. If significant relationship between social media activities and abnormal returns can be found in top trading stocks, more evidence must be found in less active ones. Therefore, it only need to focus on highly trading stocks which simultaneously have sufficient discussions in social media. Stocks in the sample are selected based on the daily average trading volume and social media attention from October 2013 to March 2014.

3.3 Sample Analysis and Determination of Time Window Size

To have a better understanding about the characteristics of social media activities, the weighted average influence (WAI) of user ***i*** is defined on its direct neighbors as follows:

$$WAI_i = \overline{G(\imath)} = \frac{1}{\|N_i\|} \sum_{i \in N_i} G(i) \tag{17}$$

The value of WAI is between 0 and 1. The higher WAI, the larger influence ***i*** has on each of its neighbor. If we set 0.5 as the baseline to separate "high WAI" and "low WAI", **Figure 8** shows one example of the percentages of two groups of users for 10 stocks out of all 34 selected ones in all sample, which are also mostly discussed in this section's analysis. As can be seen, the number of high WAI users is higher than the number of low WAI users for most of stocks, while stock 000709 and 600221 are two exceptions.

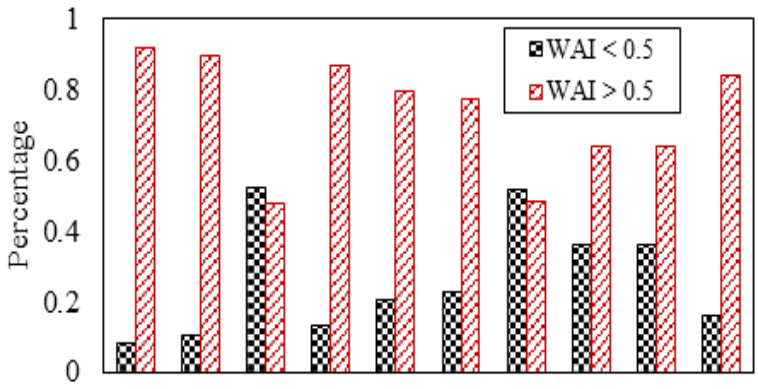


Figure 8. Percentage of the Weighted Average Influence (WAI)

*From left to right: Stock 000100, 000157, 000709, 000783, 002024, 600048, 600221, 600688, 601018, 601989

To determine appropriate moving window size for the influence updating, we check the frequency of new social media posts. Based on our data, the average frequency of articles being posted by the users is about 7 days and the average time for all these users with at least one post is about 28 days. Therefore, the influence is updated for every 7 days (1 week) based on the information from social media in the past 28 days (4 weeks).

3.4 Regression Models and Experimental results

The initial evaluation is to answer the key question: what type of market movements does the DSA framework really captured? The return or the abnormal return. Considering that uninformed traders are in the majority, the effectiveness of the framework is validated by testing the *Hypothesis 1* and *Hypothesis 2* formulized in section 3.1.

Regressions models are designed to test the two hypotheses formalized before. *Hypothesis 1* suggests that the DSA cannot directly capture the stock return if the major social media users are uninformed market traders. This hypothesis is tested by running a cross-sectional regression by model (i) as follow:

$$r_{q,t} = \delta_0 + \delta_1 dsa_{q,t} + \delta_2 r_{q,t-1} + \epsilon_{q,t} \quad (18)$$

where $r_{q,t}$ is the log-return of stock q at time t; $dsa_{q,t}$ is the logarithmic transformation for DSA of stock q at time t; δ_0 is a constant term and δ_1 is the DSA coefficient; δ_2 is the coefficient of one lag return, $r_{q,t-1}$; $\epsilon_{q,t}$~i.i.d is the error term. DSA are non-negative because social influence is positive all the time, while returns can be positive or negative. Although DSA cannot separate optimistic and pessimistic discussions in social media, the return with one lag is included as an independent variable to capture momentum of the stock price movement. Following *Hypothesis 1*, no significant relationship between the return and DSA should be found if uninformed market players dominate the social media activities.

Based on our *Hypothesis 2*, it expects to find positive relationship between the absolute value of abnormal return and DSA; hence model (ii) formulized as follows:

$$ar_{q,t} = \delta_0' + \delta_1' dsa_{q,t} + \epsilon_{q,t}' \quad (19)$$

where $ar_{q,t} = \frac{1}{2} log(AR_{q,t}{}^2)$ is the logarithmic transformation for abnormal return of stock q at time t; δ_0' is a constant term and δ_1' is the DSA coefficient; $\epsilon_{q,t}'$~i.i.d is the error term.

As an additional check, the relationship between the trading volume and DSA is verified by model (iii) as follows:

$$vol_{q,t} = \delta_0'' + \delta_1'' dsa_{q,t} + \epsilon_{q,t}'' \quad (20)$$

where $vol_{q,t}$ is the logarithmic transformation for trading volume of stock q at time t; δ_0'' is a constant term and δ_1'' is the coefficient between $vol_{q,t}$ and $dsa_{q,t}$; $\epsilon_{q,t}'' \sim i.i.d$. A positive relationship is expected to be found between the volume and DSA.

The DSA is computed based on both the market-based approach and the stock-based approach. All series are confirmed to be stationary based on the ADF test and shown the results in **Table 5**. For the estimators in above three models, student's t test is conducted to verify the significance. A p-value of 0.10 indicates a 90 percent confidence level.

Table 5. Stationarity Tests Results

Variables	ADF Test (p-value)
$r_{q,t}$	0.0120
$ar_{q,t}$	0.0103
$vol_{q,t}$	0.0228
$dsa_{q,t}$	0.0191

Table 6 reports the results of three testing models based on the DSA computed by the market-based approach. Several empirical findings can be concluded from the results. First, among the 10 stocks in sample, no relationship between the stock return and DSA can be found based on the results of model (i). None of the p-values of DSA coefficients is less than 0.1. The result is consistent with our *Hypothesis 1*. Second, positive relationship between the absolute value of abnormal return and DSA is identified for 8 stocks in 10 based on model (ii). The two exceptions are stock 000709 and stock 600221. The first one has a negative DSA coefficient which is meaningless in terms of economic implications, so the insignificant result (p-value = 0.8982) is still as expected; the second one has a positive coefficient with p-value = 0.6730. Another interesting finding is that the two stocks happen to be the two anomalies in terms of the distributions of WAI (see **Figure 8**). Therefore, the results are in favor of *Hypothesis 2*. Third, the results of model (iii) show significant relationship between the trading volume and DSA for all stocks in sample.

Table 6. Significance Tests based on Market-based Approach

SZSE Market								
Stock Code		(i) Log-Return ($r_{q,t}$)			(ii) Logarithmic Abnormal Return ($ar_{q,t}$)		(iii) Logarithmic Volume ($vol_{q,t}$)	
		(Intercept) (e-04)	$dsa_{q,t}$ (e-07)	$r_{q,t-1}$	(Intercept)	$dsa_{q,t}$	(Intercept)	$dsa_{q,t}$
000100	Coefficient	-1.1100	-6.5460	-0.2388	-6.2710	0.0401	15.5344	0.1571
	p-value	(0.1398)	(0.8456)	(0.0000)	(0.0000)	(0.0258)	(0.0000)	(0.0000)
000157	Coefficient	-2.0130	-9.1670	-0.1931	-6.8139	0.0419	13.8218	0.1208
	p-value	(0.0000)	(0.5100)	(0.0000)	(0.0000)	(0.0360)	(0.0000)	(0.0000)
000709	Coefficient	-0.8442	0.3309	-0.3350	-6.7208	-0.0606	13.6853	0.4416
	p-value	(0.2702)	(0.9903)	(0.0000)	(0.0000)	(0.8982)	(0.0000)	(0.0000)
000783	Coefficient	-3.324	2.2690	-0.1127	-6.4017	0.1114	13.8036	0.2395
	p-value	(0.0000)	(0.1820)	(0.0000)	(0.0000)	(0.0074)	(0.0000)	(0.0000)
002024	Coefficient	-2.400	-9.4970	-0.1414	-6.3772	0.1249	15.1372	0.3618
	p-value	(0.0176)	(0.6861)	(0.0000)	(0.0000)	(0.0221)	(0.0000)	(0.0000)

Table 6. Significance Tests based on Market-based Approach *(Cont.)*

SSE Market								
Stock Code		(i) Log-Return ($r_{q,t}$)			(ii) Logarithmic Abnormal Return ($ar_{q,t}$)		(iii) Logarithmic Volume ($vol_{q,t}$)	
		(Intercept) (e-04)	$dsa_{q,t}$ (e-07)	$r_{q,t-1}$	(Intercept)	$dsa_{q,t}$	(Intercept)	$dsa_{q,t}$
600048	Coefficient	-3.0760	-1.1280	-0.1182	-6.3928	0.0296	14.3525	0.1091
	p-value	(0.0000)	(0.8182)	(0.0000)	(0.0000)	(0.0483)	(0.0000)	(0.0000)
600221	Coefficient	-4.4840	0.3089	-0.3872	-6.8367	0.0490	13.9949	0.0860
	p-value	(0.5454)	(0.9250)	(0.0000)	(0.0000)	(0.6730)	(0.0000)	(0.0000)
600688	Coefficient	-3.2060	11.0800	-0.1982	-6.2686	0.0480	14.0461	0.8334
	p-value	(0.0000)	(0.7493)	(0.0000)	(0.0000)	(0.0395)	(0.0000)	(0.0000)
601018	Coefficient	-1.5230	-2.3390	-0.2892	-6.7546	0.0204	13.3157	0.4356
	p-value	(0.0285)	(0.1158)	(0.0000)	(0.0000)	(0.0174)	(0.0000)	(0.0000)
601989	Coefficient	-2.9620	8.2880	-0.1616	-6.6077	0.1152	14.8642	0.3992
	p-value	(0.0000)	(0.9278)	(0.0000)	(0.0000)	(0.0069)	(0.0000)	(0.0000)

Table 7. Significance Tests based on Stock-based Approach

SZSE Market								
Stock Code		(i) Log-Return ($r_{q,t}$)			(ii) Logarithmic Abnormal Return ($ar_{q,t}$)		(iii) Logarithmic Volume ($vol_{q,t}$)	
		(Intercept) (e-04)	$dsa_{q,t}$ (e-07)	$r_{q,t-1}$	(Intercept)	$dsa_{q,t}$	(Intercept)	$dsa_{q,t}$
000100	Coefficient	-1.0150	-9.0820	-0.2388	-6.2636	0.0153	15.5338	0.1647
	p-value	(0.1781)	(0.2534)	(0.0000)	(0.0000)	(0.0729)	(0.0000)	(0.0000)
000157	Coefficient	-2.0380	-9.048	-0.1930	-6.8146	0.0202	13.8214	0.1201
	p-value	(0.0000)	(0.3960)	(0.0000)	(0.0000)	(0.0312)	(0.0000)	(0.0000)
000709	Coefficient	-0.9334	8.0400	-0.3352	-6.3988	-0.0717	13.6856	0.4359
	p-value	(0.2434)	(0.6999)	(0.0000)	(0.0000)	(0.8168)	(0.0000)	(0.0000)
000783	Coefficient	-3.0650	-15.3500	-0.1124	-6.7218	0.1164	13.8029	0.2365
	p-value	(0.0000)	(0.1407)	(0.0000)	(0.0000)	(0.0134)	(0.0000)	(0.0000)
002024	Coefficient	-2.6360	3.4260	-0.1417	-6.4088	0.1295	15.1363	0.3612
	p-value	(0.0148)	(0.7118)	(0.0000)	(0.0000)	(0.0094)	(0.0000)	(0.0000)

Table 7. Significance Tests based on Stock-based Approach *(Cont.)*

SSE Market								
Stock Code		(i) Log-Return ($r_{q,t}$)			(ii) Logarithmic Abnormal Return ($ar_{q,t}$)		(iii) Logarithmic Volume ($vol_{q,t}$)	
		(Intercept) (e-04)	$dsa_{q,t}$ (e-07)	$r_{q,t-1}$	(Intercept)	$dsa_{q,t}$	(Intercept)	$dsa_{q,t}$
600048	Coefficient	-3.1550	1.3260	-0.1183	-6.8347	0.0340	14.3507	0.1137
	p-value	(0.0000)	(0.8556)	(0.0000)	(0.0000)	(0.0773)	(0.0000)	(0.0000)
600221	Coefficient	-3.9500	-1.0100	-0.3872	-6.7561	0.0486	13.9930	0.0715
	p-value	(0.5872)	(0.8567)	(0.0000)	(0.0000)	(0.6301)	(0.0000)	(0.0036)
600688	Coefficient	-3.1800	4.8480	-0.1983	-6.2673	0.0459	14.0445	0.8258
	p-value	(0.0000)	(0.9015)	(0.0000)	(0.0000)	(0.0578)	(0.0000)	(0.0000)
601018	Coefficient	-1.5110	-2.4750	-0.2894	-6.3930	0.0301	13.3156	0.4406
	p-value	(0.0298)	(0.9770)	(0.0000)	(0.0000)	(0.0451)	(0.0000)	(0.0000)
601989	Coefficient	-2.7560	-3.8750	-0.1602	-6.6110	0.1198	14.8640	0.4055
	p-value	(0.0000)	(0.2574)	(0.0000)	(0.0000)	(0.0119)	(0.0000)	(0.0000)

Table 7 reports the results of same models based on the DSA computed by the stock-based approach. The results are consistent with the conclusions made for the market-based approach. First, *Hypothesis 1* is supported by the results of model (i), in which there is no evidence to show the relationship between the stock return and DSA. Second, significant evidence is found to support the positive relationship between magnitude of the abnormal return and DSA for 8 of 10 stocks in the sample. Also, stock 000709 and stock 600221 are still the two exceptions. Hence, *Hypothesis 2* is supported.

In summary, the effectiveness of our DSA framework is verified for both of the market-based approach and the stock-based approach. It concludes that the DSA is significantly correlated to the absolute abnormal return and trading volume, while it does not directly affect the stock return. The testing results suggest that DSA serves as an important factor to link social media activities and the stock market. It will contribute on research and practice in finance, such as price forecasting, risk management, and other asset pricing problems.

Chapter 4 Weighted DSA and Price Shocks Detection

In the previous chapter, a degree of social attention (DSA) framework with a purpose of stock market analysis is introduced and tested. In this chapter, the historical DSA is considered and its time effect as well, in order to improve this DSA measurement. Same data sample is performed in this analysis while the stock data is expanded to the full sample size. Also, it conducts more comprehensive validation by testing the full sample as well as several subgroups separated by business sectors and listing markets.

The weighted DSA is presented as well as the influence propagation function with time effects [41]. And the model is evaluated with two major considerations as follows: 1) investigating whether the DSA, as the major index proposed in this work, has significant association with abnormal returns; 2) checking whether the computational cost is affordable during the influence updating process.

4.1 Weighted DSA and Influence Propagation Function

For a time point $t > t_e$, a weight, $\eta_q(t)$ is assigned to measure the time effect of the current DSA for stock q. Then, the weighted degree of social attention (WDSA) for stock q at time t, denoted by $\omega_q(t)$, can be expressed as:

$$\omega_q(t) = \eta_q(t)\mathcal{A}_q(t_s \to t_e) \tag{21}$$

where $\eta_q(t)$ describes the time effect of $\mathcal{A}_q(t_s \to t_e)$ at t. Let $t_f > t_i > t_e > t_s$, the accumulative WDSA is further defined for a time interval $[t_i, t_f]$ as:

$$\mathcal{W}_q|_{t_i}^{t_f} = \int_{t_i}^{t_f} \omega_q(t)\, dt = \mathcal{A}_q(t_s \to t_e) \int_{t_i}^{t_f} \eta_q(t)\, dt \tag{22}$$

where $\int_{t_i}^{t_f} \eta_q(t)\, dt$ is the periodic time effect for $[t_i, t_f]$, and $\eta_q(t)$ measures the time-varying influence propagation speed.

Assuming influence propagation follows a jump-diffusion process. That is, after a new information flow is generated by a social network user, its market influence jumps to the peak in a relatively short time, and then it slowly converges to zero. This process can be modeled by using the Gamma distribution for the flexibility in fitting different shapes of jump-diffusion. Given a shape parameter κ and an inverse scale parameter τ, the propagation at time t can be written as follows:

$$\eta(t) = \frac{1}{\Gamma(\alpha)\tau^{\kappa}} t^{\kappa-1} e^{-\frac{t}{\tau}} \tag{23}$$

where $\Gamma(\alpha)$ is the gamma function evaluated at α. The cumulative density function of **Equation (23)** can be written as:

$$\mathrm{H}(t) = \int_0^t \eta(x)dx = \frac{1}{\Gamma(\kappa)}\gamma\left(\kappa, \frac{t}{\tau}\right) \tag{24}$$

where $\gamma\left(\kappa, \frac{t}{\tau}\right)$ is the lower incomplete gamma function. The initial value of the propagation density function $\eta(t)$ is set to zero, which indicates that a new information has no influence to the social network before it is propagated. For the purpose of implementation, $\int_{t_i}^{t_f} \eta(t)$ is approximately measured by the sum of the social influence of all users who response to all related new information during $[t_i, t_f]$.

To match the current DSA defined by **Equation (11)** in Chapter 2, the time effect during $[t_s, t_e]$ is normalized to one, and **Equation (22)** can be rewritten as follows:

$$\widehat{\mathcal{W}}_q(t_i \to t_f) = \mathcal{A}_q(t_s \to t_e)\frac{\int_{t_i}^{t_f} \eta_q(t)\,dt}{\int_{t_s}^{t_e} \eta_q(t)\,dt} \tag{25}$$

where $\widehat{\mathcal{W}}_q(t_i \to t_f)$ is the normalized WDSA in $[t_i, t_f]$. And the WDSA for $[t_s, t_e]$ is normalized as $\widehat{\mathcal{W}}_q(t_i \to t_f) = \mathcal{A}_q(t_s \to t_e)$. In **Figure 9**, it shows how the influence of DSA for $[t_s, t_e]$ is propagated to $[t_i, t_f]$.

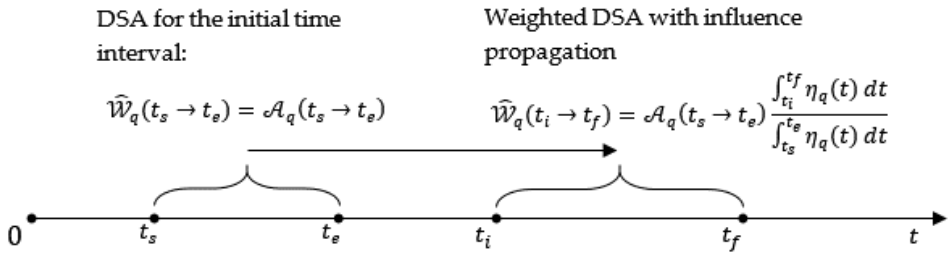
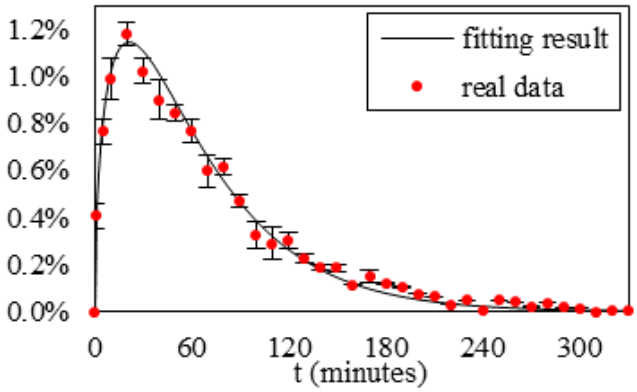
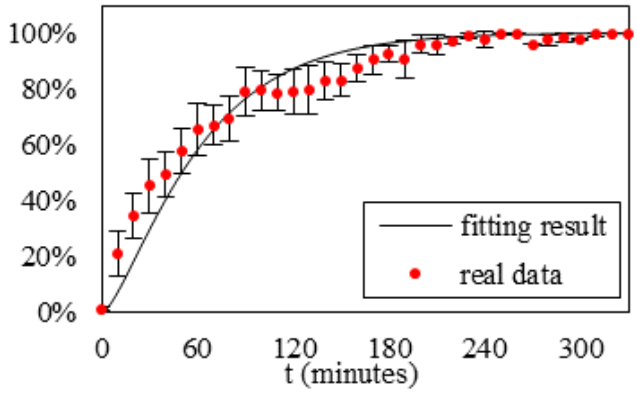


Figure 9. Influence Propagation of DSA

Figure 10 plots the estimation of $\eta(t)$ and $\mathrm{H}(t)$ based on the dynamic social influence of discussion participants of selected hot articles, which are defined as those with the number of reposts over 100. As shown in **Figure 10(a)**, within each 10-minute time window, it plots the social influence by red dots with error bars. Considering that influence propagation has a high speed in the early stage, the data of the first five minutes are plotted based on one-minute frequency for better presentation, and then 10-minute data are used for the rest of time. Also, given a time point T_m, $\eta(t) \to 0$ for $t \to T_m$, it considers that the cumulative influence of a new information flow is approximately fully propagated in a social network in $[0, T_m]$; hence, the cumulative influence (red dots) for the time period $[0, T_m]$ is normalized to 100%, see **Figure 10(b)**. In this study, since one-day trading period in the Chinese stock market is 5.5 hours, $T_m = 330$ (minutes) is set. Furthermore, the influence and cumulative data are fitted based on **Equation (23)** and **Equation (24)**, and the solid lines are the fitting results with $\kappa = 1.4958$ and $\tau = 42.2441$. The real data is well fitted by the proposed propagation function.



(a) Propagation density function $\eta(t) = \frac{1}{\Gamma(\alpha)\tau^{\kappa}} t^{\kappa-1} e^{-\frac{t}{\tau}}$



(b) Cumulative distribution function $\mathrm{H}(t) = \int_0^t \eta(x)dx = \frac{1}{\Gamma(\kappa)} \gamma\left(\kappa, \frac{t}{\tau}\right)$

Figure 10. Propagation Density Function and Cumulative Distribution Function (Parameter Estimations: $\kappa = 1.4958$ $\tau = 42.2441$).

The time effects of DSA are estimated based on selected hot articles (with the number of discussion participators more than 100) from our data using propagation functions **Equation (23)** and **Equation (24)**. To test effects of DSA over different time, the cumulative weights are separated into five time intervals by time points, $t_m \in \{t_0, t_1, t_2, t_3, t_4, t_5\}$ where $t_0 = 0$ and $t_5 = 330$ indicate the start and ending time index of one-day trading period respectively. As shown in **Figure 11**, we set $t_1 = 10$, and the weight $\eta_q(t_1) = 1$. Then, the weighted current DSA is denoted by $wdsa_{q,t_1} = \log\big(\mathcal{W}_q(t \to t + t_1)\big) = \log(\mathcal{A}_q\ (t \to t + t_1))$. t_2 is set to be 30, the time point when the weight reaches the maximum value ($\eta_q(t_2) = \eta_{max}$); we set $t_3 = 100$ when the weight drops to about 50% of the initial weight ($\eta_q(t_3) = 0.5$); t_4 is set to be 180, the time point when the weight drops to 10% of its initial value ($\eta_q(t_4) = 0.1$). In the following sections, the three-factor form model with $T = 2$ is tested as well as the full model which has $T = 4$.

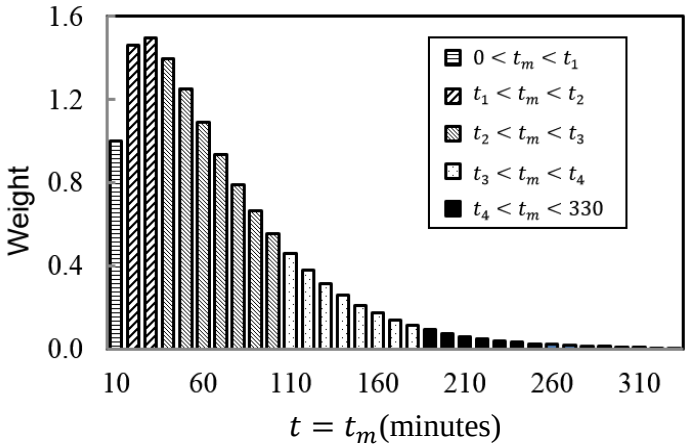


Figure 11. Ten-minute Cumulative Weights

4.2 DSA, Returns and Abnormal Returns

According to Section 3.4, regressions models are designed to test the two hypotheses formalized before. *Hypothesis 1* is tested by running a cross-sectional regression model as **Equation (18)** and *Hypothesis 2* is tested by running a one-factor model as **Equation (19)**.

These models **Equation (18)** and **Equation (19)** are considered as the basic models of DSA which only consider the effect of current DSA. To improve the models, the relationship between historical DSAs and abnormal returns is further investigated. Let $t_m \in \{t_0, t_1, t_2, \dots, t_T\}$ and $t_0 = 0$, it is set:

$$wdsa_{q,t-t_m} = \log(\widehat{\mathcal{W}}_q(t - t_{m-1} \to t - t_m)) \tag{26}$$

and then an extended version of **Equation (19)** is formulized as follows:

$$ar_{q,t} = \zeta + \phi_0 dsa_{q,t} + \sum_{m=1}^{T} \phi_m wdsa_{q,t-t_m} + u_{q,t} \tag{27}$$

where $T > 0$; ζ is a constant term and ϕ_m is the coefficient of $wdsa_{q,t_m}$; $u_{q,t}$~i.i.d. is the error term. Based on this model, the time effect of DSA on stock abnormal returns is investigated. Moreover, we expect to obtain the best fitting performance from the full model in which T is the possible maximum.

Based on *Hypothesis 2*, positive relationships are expected to be found between the absolute value of abnormal returns and the current DSA, $dsa_{q,t}$, as well as some weighted DSA, $wdsa_{q,t-t_m}$, when m is small. The DSA is computed by the stock-based approach. For all regression models, student's t tests are conducted to check the significance of each estimator. A p-value less than 0.1 indicate the estimator is within a 90% confidence level; a p-value less than 0.05 indicate the estimator is within a 95% confidence level. It is also compared the adjusted R^2 for differernt models to study the improvement of goodness-of-fit.

4.2.1 Current DSA Analysis

The effects of current DSA on stock returns and abnormal returns based on model **Equation (18)** and model **Equation (19)** is studied respectively. In addition to the test based on full sample, the data are separated into several groups based on two methods: business sectors and listing markets. **Table 8** reports the estimated coefficients and p-values of the two models based on nine samples. Panel A reports the results based on six subsamples categorized by different business sectors. Panel B reports the test results based on the full sample and two subsamples categorized by two listing markets (SSE and SZSE). Several empirical findings can be concluded from the results. First, no evidence can be found to support the relationship between the stock return and the current DSA. It can be seen that all p-values of DSA coefficient are too large to claim significant findings for all samples. The smallest p-values are in the "service" subsample (p-value = 0.1454) and the "utility" subsample (p-value = 0.1436). These results support our

Hypothesis 1 that suggests that there is no direct association between social media activities and stork returns if uninformed traders dominate discussions in a social network. Second, evidence is found to support the positive relationship between the absolute value of abnormal return and the current DSA. As shown in the **Table 8**, results from most of the samples provide evidence to support this conclusion. An exception is found from the result for one business sector: industrial goods. Besides that, all other results are significant at the 99% confidence level. Therefore, it believes that *Hypothesis 2* is supported. However, the fitting performance is not good for basic models. The adjusted R^2 is reported for model **Equation (19)**, and it can be seen that they are generally less than 10%. The minimum one is from the subsample of consumer goods that has its adjusted R^2 equal to 0.0760 (only 7.60% of abnormal returns are explained by the current DSA).

4.2.2 Weighted DSA Analysis

To investigate the time effect of previous social media activities, model **Equation (27)** is tested with different forms based on the nine samples tested above. **Table 9** reports results from the three-factor form model ($T = 2$) in which the weights for historical DSAs are estimated based on selected article samples. Panel A reports the results based on six subsamples categorized by different business sectors. Panel B reports the test results based on the full sample and two subsamples categorized by two listing markets (SSE and SZSE). Especially, the afternoon session data is used only in the empirical tests to guarantee the existence of $wdsa_{t-t_1}$ and

$wdsa_{t-t_2}$ for all observations, hence fair tests are conducted. As shown, it first confirms that the current DSA, dsa_t, is significantly correlated to stock abnormal returns for all samples. To be specific, dsa_t in the full sample and all the subsamples is found to be significant at the 99% confidence level. Furthermore, the two newly included weighted DSAs are found to be significantly correlated to stock abnormal returns as well. Among the results of $wdsa_{t-t_1}$ from the nine samples, it gets five at the 95% level, and four in 90% level. For $wdsa_{t-t_2}$, six are at the 95% level, and three are at the 90% level. These results are also consistent with *Hypothesis 2*. The adjusted R^2 is then reported for each sample. Comparing to the basic model, it can be seen significant improvements in terms of the performance of data fitting. The minimum adjusted R^2 is found from the subsample of foods (U) that has an adjusted R^2 equals to 0.1600; the average adjusted R^2 is about 0.1727.

Table 10 reports the results from the full model ($T = 4$). Similar to the three factor form, to avoid missing weighted DSAs for all observations, the full model is tested where the weighted DSAs are estimated based on the minute level data, while the tested abnormal returns are counted once per day. Panel A reports the results based on six subsamples categorized by different business sectors. Panel B reports the test results based on the full sample and two subsamples categorized by two listing markets (SSE and SZSE). As shown, first, it identifies significant relationship between the stock abnormal returns and the current DSA. Significant evidence is also found to support same relationship for $wdsa_{t-t_1}$ and $wdsa_{t-t_2}$. These findings are consistent with *Hypothesis 2*. However, it cannot find evidence to support the relationship for

$wdsa_3$ from some samples, such as the whole market, SZSE and some business sectors like BM, CG, F, and IG. Similar results can be found for $wdsa_{t-t_4}$ as well. These results show that the effect of DSA on the stock market decreases while time passes. More importantly, it is found that the adjusted R^2 is greater than that of the basic model and three-factor model. The average adjusted R^2 is about 0.2657, and the minimum value is 0.2586 which is obtained from the listing market SZSE.

Test results from three different models sufficiently support the *Hypothesis 1* and *Hypothesis 2* which are formalized based on a widely accepted assumption that suggests the Chinese stock market is semi-strong inefficient. Hence, the findings also support the semi-strong inefficiency assumption. On the other hand, the consistent findings further confirm the effectiveness of our DSA framework and the time effect estimating process. To compare the fitting performance of the basic model, the three-factor model, and the full model. **Figure 12** shows their adjusted R^2 based on different samples. The results clearly show the dominant of the full model in fitting real world data. Values of the adjusted R^2 resuled from the full model are at least doubled, comparing to the basic model, even if our results show that the significance of DSA effects starts reducing when weights drop to 50% of the initial value. Furthermore, improvements can also be seen from the three-factor form model to the full model.

Table 8. Significance Test for the Current DSA (Basic Models)

Panel A						
Business Sectors	Model 1: Log-Return ($r_{q,t}$)			Model 2: Logarithmic Abnormal Return ($ar_{q,t}$)		
	Intercept	$dsa_{q,t}$	$r_{q,t-1}$	Intercept	$dsa_{q,t}$	Adjusted R^2
BM	0.0021	0.0003	0.1750	-6.4548	0.1153***	0.0921
	(0.0000)	(0.5913)	(0.0000)	(0.0000)	(0.0000)	
CG	0.0025	0.0003	0.1110	-6.3987	0.1177***	0.0900
	(0.0000)	(0.2946)	(0.0000)	(0.0000)	(0.0000)	
F	0.0022	0.0003	0.1360	-6.5000	0.1085***	0.0760
	(0.0000)	(0.6169)	(0.0000)	(0.0000)	(0.0000)	
IG	0.0022	0.0001	0.1280	-6.4870	0.1005***	0.0850
	(0.0000)	(0.2951)	(0.0000)	(0.0000)	(0.0000)	
S	0.0020	0.0002	0.1380	-6.4660	0.1039***	0.0936
	(0.0000)	(0.1454)	(0.0000)	(0.0000)	(0.0000)	
U	0.0022	0.0003	0.1150	-6.4001	0.1045***	0.1005
	(0.0000)	(0.1436)	(0.0000)	(0.0000)	(0.0000)	

Table 8. Significance Test for the Current DSA (Basic Models) *(Cont.)*

Panel B						
Listing Markets	Model 1: Log-Return ($r_{q,t}$)			Model 2: Logarithmic Abnormal Return ($ar_{q,t}$)		
	Intercept	$dsa_{q,t}$	$r_{q,t-1}$	Intercept	$dsa_{q,t}$	Adjusted R^2
SSE	0.0022	0.0002	0.1440	-6.4960	0.1029***	0.0810
	(0.0000)	(0.9517)	(0.0000)	(0.0000)	(0.0000)	
SZSE	0.0023	0.0003	0.1490	-6.4524	0.1098***	0.1040
	(0.0000)	(0.6595)	(0.0000)	(0.0000)	(0.0000)	
Full Sample	0.0022	0.0003	0.1470	-6.4739	0.1196***	0.0840
	(0.0000)	(0.4903)	(0.0000)	(0.0000)	(0.0000)	

Notes: This table reports the estimated coefficients of the two basic models.

Model 1: $r_{q,t} = \delta_0 + \delta_1 dsa_{q,t} + \delta_2 r_{q,t-1} + \epsilon_{q,t}$ *and Model 2:* $ar_{q,t} = \delta_0' + \delta_1' dsa_{q,t} + \epsilon_{q,t}'$.

*P-values are reported in parentheses. * Statistically significant at the 90% confidence level; ** Statistically significant at the 95% confidence level; *** Statistically significant at the 99% confidence level.*

Table 9. Significance Test for the Weighted DSA (The Three-Factor Form)

Panel A					
Business Sectors	Intercept	dsa_t	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	Adjusted R^2
BM	-6.5567 (0.0000)	0.1295*** (0.0000)	0.0086** (0.0331)	0.0135* (0.0822)	0.1720
CG	-6.5439 (0.0000)	0.1497*** (0.0000)	0.0138* (0.0954)	0.0149* (0.0516)	0.1630
F	-6.6018 (0.0000)	0.1180*** (0.0000)	0.0133** (0.0243)	0.0129** (0.0279)	0.1731
IG	-6.5599 (0.0000)	0.0975*** (0.0000)	0.0108** (0.0473)	0.0127** (0.0289)	0.1635
S	-6.5096 (0.0000)	0.1124*** (0.0000)	0.0105* (0.0752)	0.0127** (0.0124)	0.1740
U	-6.5293 (0.0000)	0.1432*** (0.0000)	0.0151* (0.0679)	0.0155** (0.0404)	0.1600

Table 9. Significance Test for the Weighted DSA (The Three-Factor Form) *(Cont.)*

Panel B					
Listing Markets	Intercept	dsa_t	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	Adjusted R^2
SSE	-6.5516 (0.0000)	0.1550*** (0.0000)	0.0114** (0.0242)	0.0137* (0.0715)	0.1940
SZSE	-6.6028 (0.0000)	0.1069*** (0.0000)	0.0106** (0.0292)	0.0164** (0.0125)	0.1721
Full Sample	-6.5772 (0.0000)	0.1404*** (0.0000)	0.0103* (0.0854)	0.0092** (0.0211)	0.1830

Notes: This table reports the estimated coefficients of the three-factor model:

$$ar_{q,t} = \zeta + \phi_0 dsa_{q,t} + \sum_{m=1}^{2} \phi_m wdsa_{q,t-t_m} + u_{q,t}.$$

*P-values are reported in parentheses. * Statistically significant at the 90% confidence level; ** Statistically significant at the 95% confidence level; *** Statistically significant at the 99% confidence level.*

Table 10. Significance Test for the Weighted DSA (Full Model)

Panel A							
Business Sectors	Intercept	dsa_t	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	$wdsa_{t-t_3}$	$wdsa_{t-t_4}$	Adjusted R^2
BM	-6.4654 (0.0000)	0.1230*** (0.0000)	0.0120* (0.0739)	0.0159*** (0.0050)	0.0059 (0.1890)	0.0021 (0.4223)	0.2686
CG	-6.6151 (0.0000)	0.1060*** (0.0000)	0.0146** (0.0114)	0.0149*** (0.0022)	-0.0011 (0.2252)	-0.0011** (0.0108)	0.2631
F	-6.5233 (0.0000)	0.1370*** (0.0000)	0.0137** (0.0227)	0.0131* (0.0548)	0.0033 (0.9232)	-0.0007* (0.0908)	0.2790
IG	-6.5230 (0.0000)	0.0888*** (0.0000)	0.0206*** (0.0074)	0.0117** (0.0223)	0.0016 (0.2251)	-0.0004 (0.2794)	0.2665
S	-6.5640 (0.0000)	0.1191*** (0.0000)	0.0119** (0.0368)	0.0130** (0.0284)	-0.0037* (0.0563)	-0.0012** (0.0296)	0.2573
U	-6.5610 (0.0000)	0.1530*** (0.0000)	0.0140** (0.0294)	0.0135*** (0.0020)	0.0029*** (0.0029)	-0.0002 (0.5372)	0.2613

Table 10. Significance Test for the Weighted DSA (Full Model) *(Cont.)*

Panel B							
Listing Markets	Intercept	dsa_t	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	$wdsa_{t-t_3}$	$wdsa_{t-t_4}$	Adjusted R^2
SSE	-6.4850	0.1171***	0.0127**	0.0130**	-0.0012**	-0.0010**	0.2709
	(0.0000)	(0.0000)	(0.0371)	(0.0108)	(0.0318)	(0.0329)	
SZSE	-6.6360	0.1400***	0.0130**	0.0147***	0.0077	-0.0033	0.2586
	(0.0000)	(0.0000)	(0.0184)	(0.0097)	(0.2893)	(0.2866)	
Full Sample	-6.5640	0.0908***	0.0125***	0.0135**	-0.0004	0.0053**	0.2659
	(0.0000)	(0.0000)	(0.0013)	(0.0267)	(0.1503)	(0.0334)	

Notes: This table reports the estimated coefficients of the three-factor model:

$$ar_{q,t} = \zeta + \phi_0 dsa_{q,t} + \sum_{m=1}^{2} \phi_m wdsa_{q,t-t_m} + u_{q,t}.$$

*P-values are reported in parentheses. * Statistically significant at the 90% confidence level; ** Statistically significant at the 95% confidence level; *** Statistically significant at the 99% confidence level.*

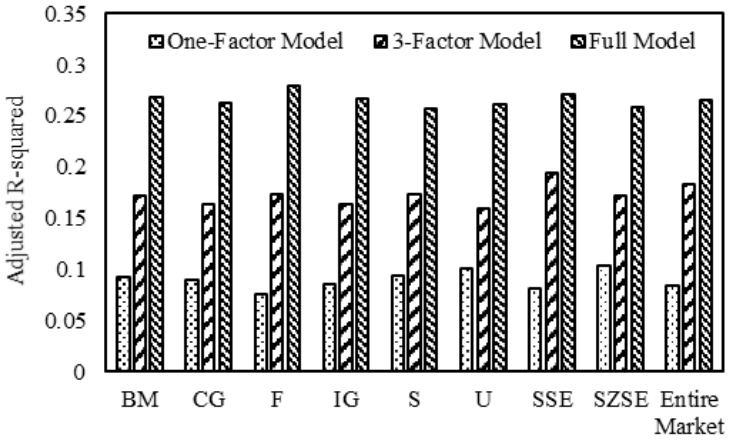


Figure 12. Comparison of the Fitting Performance for the Basic Model and Improved Models.

Especially, as an additional check, the abnormal return is separated into positive and negative ones and redo the significance tests for all the three models. The six business sectors and the total sample are selected sample size. **Table 11** lists the results come from the current DSA tests, **Table 12** lists the results come from the weighted DSA of the three-factor form and **Table 13** lists the results come from the weighted DSA of the full model.

As the results in **Table 11**, **Table 12** and **Table 13**, the consist conclusions can be made that there are significant relationship between the stock abnormal returns and the current DSA, and some forms of the weighted DSA as well. The full model gives the best fitting performance for both positive and negative abnormal return based on the adjusted R^2.

Table 11. Separated Abnormal Return Tests for the Current DSA

Sample	Positive Abnormal Return			Negative Abnormal Return		
	Intercept	$dsa_{q,t}$	Adjusted R^2	Intercept	$dsa_{q,t}$	Adjusted R^2
BM	-2.4850 (0.0000)	0.1398 (0.0000)	0.0854	-3.1730 (0.0000)	0.0765 (0.0141)	0.0920
CG	-2.4326 (0.0000)	0.1342 (0.0185)	0.0822	-3.2874 (0.0000)	0.0715 (0.0306)	0.0854
F	-2.4983 (0.0000)	0.1521 (0.0000)	0.0820	-3.3765 (0.0000)	0.0791 (0.0147)	0.0960
IG	-2.4709 (0.0000)	0.1253 (0.0123)	0.0960	-3.2857 (0.0000)	0.0761 (0.0028)	0.1142
S	-2.4748 (0.0000)	0.1072 (0.0321)	0.0905	-3.3332 (0.0000)	0.0789 (0.0000)	0.0923
U	-2.4900 (0.0000)	0.1595 (0.0000)	0.0857	-3.2310 (0.0000)	0.0703 (0.0168)	0.0863
Total	-2.4089 (0.0000)	0.1236 (0.0313)	0.0936	-3.3339 (0.0000)	0.0803 (0.0000)	0.1022

Table 12. Separated Abnormal Return Tests for the Weighted DSA (The Three-Factor Form)

Sample	Positive Abnormal Return				
	Intercept	$dsa_{q,t}$	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	Adjusted R^2
BM	-2.4116 (0.0000)	0.1335 (0.0000)	0.0247 (0.0000)	0.0093 (0.0923)	0.1787
CG	-2.4176 (0.0000)	0.1230 (0.0135)	0.0212 (0.0921)	0.0123 (0.0794)	0.1652
F	-2.4600 (0.0000)	0.1482 (0.0062)	0.0243 (0.0461)	0.0179 (0.0426)	0.1799
IG	-2.4248 (0.0000)	0.1324 (0.0325)	0.0213 (0.0001)	0.0165 (0.0785)	0.1792
S	-2.5010 (0.0000)	0.1451 (0.0000)	0.0248 (0.0000)	0.0080 (0.0135)	0.1740
U	-2.4955 (0.0000)	0.1312 (0.0000)	0.0251 (0.0038)	0.0102 (0.0899)	0.1626
Total	-2.4300 (0.0000)	0.1180 (0.0309)	0.0198 (0.0000)	0.0092 (0.1779)	0.1769

Table 12. Separated Abnormal Return Tests for the Weighted DSA (The Three-Factor Form) *(Cont.)*

Sample	Negative Abnormal Return				
	Intercept	$dsa_{q,t}$	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	Adjusted R^2
BM	-3.0815 (0.0000)	0.0766 (0.0000)	0.0191 (0.0498)	0.0096 (0.0091)	0.1834
CG	-3.2495 (0.0000)	0.0659 (0.0000)	0.0175 (0.0178)	0.0156 (0.0296)	0.1722
F	-3.2919 (0.0000)	0.0611 (0.0779)	0.0200 (0.0913)	0.0108 (0.0261)	0.1961
IG	-3.0691 (0.0000)	0.0732 (0.0222)	0.0161 (0.0483)	0.0119 (0.0790)	0.1881
S	-3.2905 (0.0000)	0.0662 (0.0814)	0.0139 (0.0000)	0.0141 (0.0199)	0.1855
U	-3.2267 (0.0000)	0.0671 (0.0157)	0.0150 (0.0892)	0.0103 (0.0178)	0.1739
Total	-3.1154 (0.0000)	0.0749 (0.0041)	0.0148 (0.0095)	0.0104 (0.0869)	0.1922

Table 13. Separated Abnormal Return Tests for the Weighted DSA (Full Model)

Sample	Positive Abnormal Return						
	Intercept	dsa_t	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	$wdsa_{t-t_3}$	$wdsa_{t-t_4}$	Adjusted R^2
BM	-2.5578 (0.0000)	0.1206 (0.0848)	0.0241 (0.0311)	0.0126 (0.0354)	0.0005 (0.0937)	0.0041 (0.5953)	0.2543
CG	-2.4207 (0.0000)	0.1056 (0.0198)	0.0295 (0.0673)	0.0120 (0.0429)	-0.0071 (0.6929)	-0.0056 (0.2091)	0.2471
F	-2.4670 (0.0000)	0.1470 (0.0000)	0.0240 (0.0317)	0.0128 (0.0599)	0.0133 (0.0000)	0.0117 (0.0043)	0.2626
IG	-2.5099 (0.0000)	0.1281 (0.0000)	0.0288 (0.0000)	0.0135 (0.0139)	0.0183 (0.5664)	0.0174 (0.0000)	0.2540
S	-2.5062 (0.0000)	0.1333 (0.0000)	0.0251 (0.0000)	0.0124 (0.0665)	-0.0207 (0.0734)	0.0285 (0.0000)	0.2590
U	-2.4904 (0.0000)	0.1233 (0.0000)	0.0284 (0.0197)	0.0138 (0.0568)	0.0037 (0.2810)	0.0130 (0.0062)	0.2416
Total	-2.5055 (0.0000)	0.1307 (0.0000)	0.0312 (0.0000)	0.0130 (0.0522)	-0.0025 (0.1275)	0.0331 (0.4999)	0.2581

Table 13. Separated Abnormal Return Tests for the Weighted DSA (Full Model) *(Cont.)*

Sample	Negative Abnormal Return						
	(Intercept)	dsa_t	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	$wdsa_{t-t_3}$	$wdsa_{t-t_4}$	Adjusted R^2
BM	-3.2868 (0.0000)	0.0714 (0.0343)	0.0160 (0.0434)	0.0113 (0.0218)	0.0065 (0.0043)	-0.0011 (0.0793)	0.2696
CG	-3.3386 (0.0000)	0.0619 (0.0280)	0.0236 (0.0016)	0.0170 (0.0388)	0.0068 (0.0000)	-0.0081 (0.1526)	0.2517
F	-3.1010 (0.0000)	0.0701 (0.0886)	0.0269 (0.0000)	0.0175 (0.0000)	0.0325 (0.0023)	0.0049 (0.5341)	0.2667
IG	-3.2128 (0.0000)	0.0844 (0.0012)	0.0320 (0.0308)	0.0112 (0.0000)	0.0433 (0.2424)	0.0313 (0.0000)	0.2641
S	-3.2651 (0.0000)	0.0980 (0.0561)	0.0203 (0.0489)	0.0103 (0.0000)	0.0536 (0.3693)	-0.0439 (0.0000)	0.2610
U	-3.2231 (0.0000)	0.0838 (0.0746)	0.0153 (0.0935)	0.0114 (0.0433)	0.0233 (0.0018)	-0.0023 (0.0044)	0.2513
Total	-3.2505 (0.0000)	0.0825 (0.0073)	0.0216 (0.0000)	0.0134 (0.0507)	-0.0021 (0.5429)	0.0277 (0.0229)	0.2630

There are also some interesting findings can be discovered in the separated abnormal return tests from **Table 11**, **Table 12** and **Table 13**. First, the coefficient for DSA of positive abnormal return is larger than the coefficient for DSA of negative abnormal return. For example in total sample of **Table 11**, the coefficient of $dsa_{q,t}$ is 0.1236 as the positive abnormal return and the coefficient of $dsa_{q,t}$ is 0.0803 as the negative abnormal return; in total sample of **Table 12**, the coefficient of $dsa_{q,t}$ is 0.1180 as the positive abnormal return and the coefficient of $dsa_{q,t}$ is 0.0749 as the negative abnormal return; in total sample of **Table 13**, the coefficient of $dsa_{q,t}$ is 0.1307 as the positive abnormal return and the coefficient of $dsa_{q,t}$ is 0.0825 as the negative abnormal return. This could be explained that the social media would be more reacted when positive abnormal return happens.

Second, the adjusted R^2 for the positive abnormal return is smaller than the one for the negative abnormal return which means the fitting performance for the negative price shocks is better than positive ones. In **Table 11**, the average adjusted R^2 is 0.0879 for the positive abnormal return and 0.0955 for the negative abnormal return; in **Table 12**, the average adjusted R^2 is 0.1738 for the positive abnormal return and 0.1845 for the negative abnormal return; in **Table 13**, the average adjusted R^2 is 0.2538 for the positive abnormal return and 0.2611 for the negative abnormal return.

Third, results for six business sectors are different with each other which means the sensitivity of price shocks for different business sector is independent. In **Table 11**, for the positive abnormal

return, the smallest coefficient of $dsa_{q,t}$ is 0.1072 of subsample "S" and the biggest is 0.1595 of subsample "U"; for the negative abnormal return, the smallest coefficient of $dsa_{q,t}$ is 0.0703 of subsample "U" and the biggest is 0.0791 of subsample "F". In **Table 12**, for the positive abnormal return, the smallest coefficient of $dsa_{q,t}$ is 0.1230 of subsample "CG" and the biggest is 0.1482 of subsample "F"; for the negative abnormal return, the smallest coefficient of $dsa_{q,t}$ is 0.0611 of subsample "F" and the biggest is 0.0766 of subsample "BM". In **Table 13**, for the positive abnormal return, the smallest coefficient of $dsa_{q,t}$ is 0.1056 of subsample "CG" and the biggest is 0.1470 of subsample "F"; for the negative abnormal return, the smallest coefficient of $dsa_{q,t}$ is 0.0619 of subsample "CG" and the biggest is 0.0980 of subsample "S".

Robust regression is also tested for the outlier control and robustness check purpose. In **Table 14**, robust regressions are performed for robust test of total sample and the results of coefficients are consistent with previous ones. These could give more reliable and effective test results.

4.3 Ranking Price Shocks

Results from the hypothesis tests statistically verified the association between the proposed DSAs and price shocks. The framework is further evaluated in a ranking problem of price shocks detection. To do so, it need to be prepared the true ranking list of price shocks based on the abnormal returns, and then compare the ranking performance of the newly proposed DSA and several benchmarks.

Table 14. Robust Regression Tests for the Total Sample

Model Form	Abnormal Return	Coefficient (P-value)					
		(Intercept)	dsa_t	$wdsa_{t-t_1}$	$wdsa_{t-t_2}$	$wdsa_{t-t_3}$	$wdsa_{t-t_4}$
Current DSA	Total	-6.5038 (0.0000)	0.1079 (0.0000)				
	Positive	-2.5492 (0.0000)	0.1067 (0.0012)				
	Negative	-3.2384 (0.0000)	0.0874 (0.0000)				
Weighted DSA (The Three-Factor From)	Total	-6.4650 (0.0000)	0.1382 (0.0000)	0.0129 (0.0198)	0.0178 (0.0275)		
	Positive	-2.5461 (0.0000)	0.1175 (0.0091)	0.0176 (0.0436)	0.0083 (0.0842)		
	Negative	-3.2349 (0.0000)	0.0718 (0.0037)	0.0142 (0.0321)	0.0127 (0.0611)		
Weighted DSA (Full Model)	Total	-6.4450 (0.0000)	0.0885 (0.0000)	0.0136 (0.0503)	0.0160 (0.0333)	-0.0008 (0.9112)	0.0087 (0.0123)
	Positive	-2.5454 (0.0000)	0.1404 (0.0008)	0.0325 (0.0222)	0.0144 (0.0783)	-0.0025 (0.5833)	0.0274 (0.0049)
	Negative	-3.2185 (0.0000)	0.0839 (0.0000)	0.0276 (0.0289)	0.0132 (0.0838)	-0.0100 (0.4431)	0.0349 (0.0329)

4.3.1 Evaluation Metrics and Benchmarks

The following evaluation methods are used in measuring of the ranking quality as they are frequently performed in articles of this field.

Normalized Discounted Cumulative Gain. The Discounted Cumulative Gain (DCG) accumulated at a particular rank position $\boldsymbol{n}$ is defined as:

$$DCG_n = \begin{cases} rel_1, & \text{if } n = 1 \\ DCG[n-1] + \dfrac{rel_n}{\log_2 n}, & \text{if } n > 1 \end{cases} \tag{28}$$

where rel_n is the graded relevance of the result at position n. In our ranking analysis, rel_n will be the stock abnormal return value. When sorting the abnormal return directly by their values, we get the maximum possible DCG till position n, also called Ideal DCG (IDCG). The normalized discounted cumulative gain (NDCG) is computed as:

$$NDCG_n = \frac{DCG_n}{IDCG_n} \tag{29}$$

all NDCG calculations are then relative values on the interval 0.0 to 1.0 and so are cross-query comparable. The larger $NDCG_n$ is, the higher top-n ranking accuracy is.

Kendall's Tau Coefficient. Kendall's Tau coefficient, or short for Tau, measures the overall ranking accuracy. Let (x_i, y_i) be a set of observations of the joint random variables X and Y repectiverly, such that all the values of x_i and y_i are unique. Any pair of observations (x_i, y_i) and (x_j, y_j), where $i \neq j$, are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant, if both $x_i > x_j$ and $y_i < y_j$ or if both $x_i < x_j$ and $y_i > y_j$. Tau is given by:

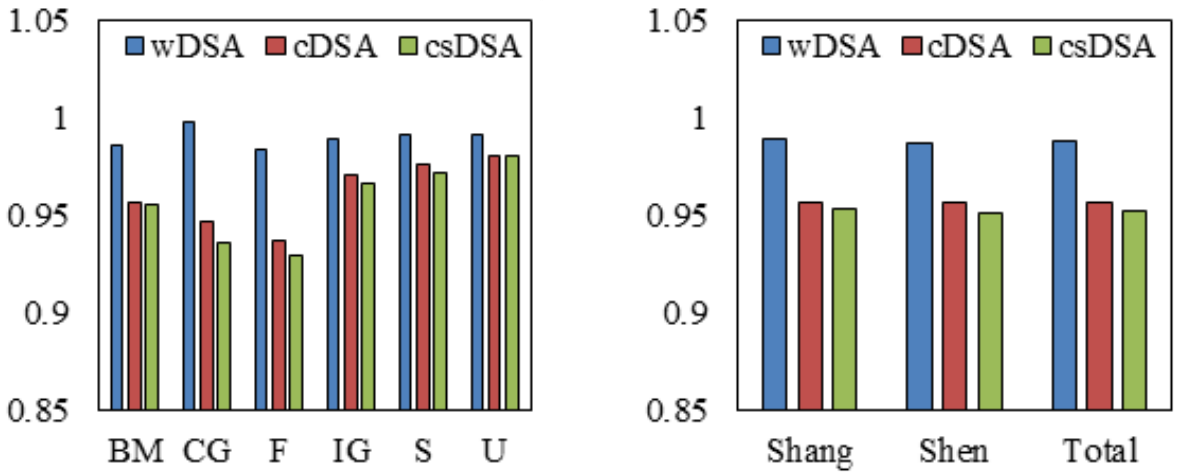
$$Tau = \frac{\#concordant - \#discordant}{\#concordant + \#discordant} \tag{30}$$

same as NDCG, the larger Tau value is, the higher ranking accuracy is.

To evaluate the newly proposed weighted DSA (wDSA), its ranking accuracy is compared against two baseline methods: the current DSA (cDSA) and the constant self-influence DSA (csDSA). cDSA solely consider the daily cumulative DSA using the current DSA value. For csDSA, the self-influence (confidence) is set equal to 1 (full confidence) and the influence of the social network is computed without updating process.

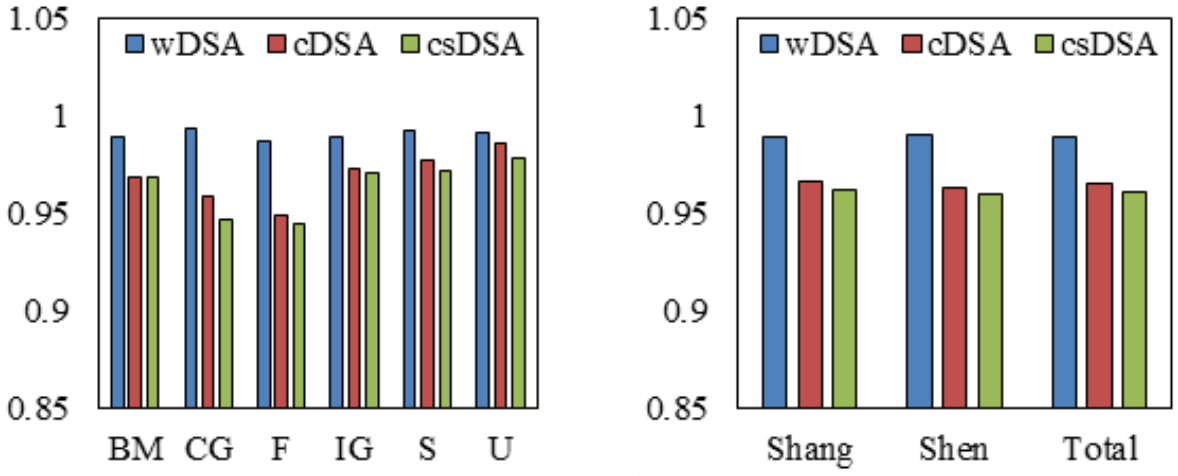
4.3.2 Overall Performance

The ranking performances are compared based on different samples including six business sectors (BM, CG, F, IG, S, U), two listing markets (Shang and Shen), and the full sample (Total).



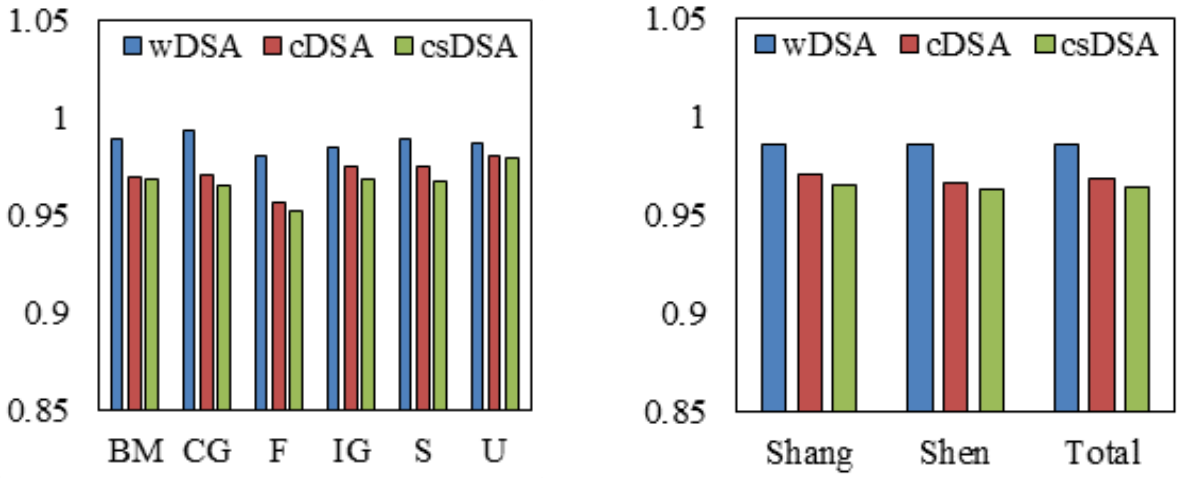
(a) Business Sectors (b) Listing Markets

Figure 13. $NDCG@n = 10$ Comparison Results



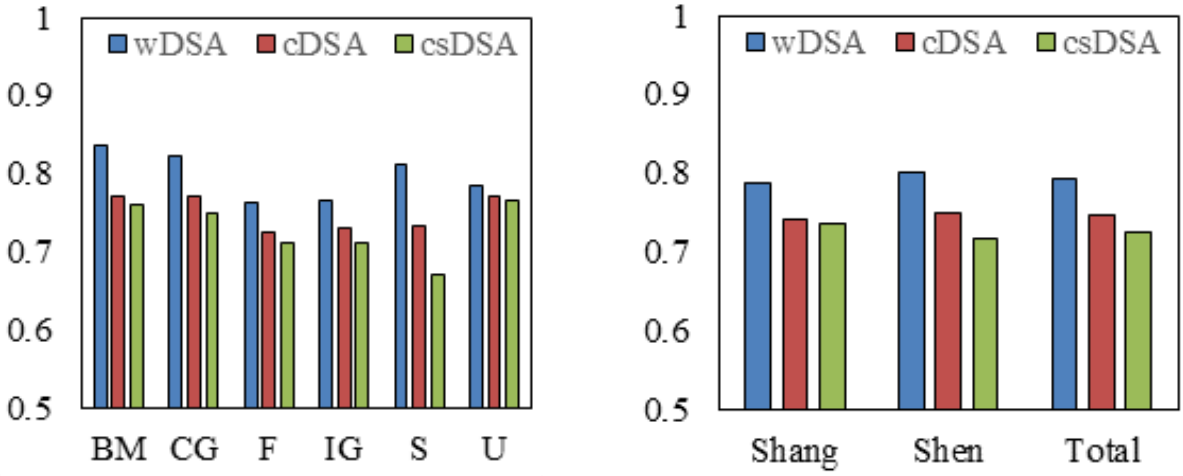
(a) Business Sectors (b) Listing Markets

Figure 14. $NDCG@n = 20$ Comparison Results



(a) Business Sectors (b) Listing Markets

Figure 15. $NDCG@n = 50$ Comparison Results



(a) Business Sectors (b) Listing Markets

Figure 16. Tau Comparison Results

Figure 13, **Figure 14** and **Figure 15** show the comparisons of NDCG in detecting daily price shocks. The position number is set $n = 10, 20, \text{and } 50$ for $NDCG_n$. **Figure 13** shows the $NDCG_{10}$ comparison results, **Figure 14** shows the $NDCG_{20}$ comparison results and **Figure 15** shows the $NDCG_{50}$ comparison results. **Figure 13(a)**, **Figure 14(a)** and **Figure 15(a)** show the NDCG comparison with different business sectors. **Figure 13(b)**, **Figure 14(b)** and **Figure 15(b)** show results with different listing markets and especially. As shown in **Figure 13**, **Figure 14** and **Figure 15**, it can be seen that our weighted DSA ranking method can achieve larger NDCG values which indicate better ranking performance.

Figure 16 shows the comparison of Tau based on different samples. **Figure 16(a)** shows the Tau comparison with different business sectors and **Figure 16(b)** shows results with different listing markets and especially. As shown in **Figure 16**, the weighted DSA ranking method achieves the largest values of Tau under all samples tested.

In summary, the above overall performance validates the effectiveness of the weighted DSA ranking methods and present the best ranking accuracy than the other two baseline algorithms.

Chapter 5 Forecasting Price Shocks with Sentiment Analysis

In this chapter, a novel framework is proposed to effectively forecast the direction of prices shock in the stock market based on stock-related social media activities and historical market information [42]. In the theory of asset pricing, a price shock or abnormal return can be defined by the difference between the actual return of an asset and the expected one. It can be considered as a market movement that cannot be explained by pricing models. Several recent studies in quantitative finance have found that capturing stock price shocks would help in solving market prediction problems [43-45]. Moreover, some studies claimed that those price shocks are correlated to investor attention [19, 46, 47]. Although it is usually difficult to model human behaviors using traditional approaches, recent innovation of the Internet technologies including social network techniques, provide alternative ways to measure the social influence of these activities quantitatively; hence their impact on financial markets can be further studied.

Figure 17 shows an example of the potential relationship between the stock abnormal return and social media activities related to Shanghai AJ Group Inc. (600643), a Chinese public firm. Here, for a quick preview, the number of opinion posts is used in Weibo (the biggest Chinese social network platform) to illustrate investors' attention. The signs of these posts are determined by the number of positive words less the number of negative words [48]. As can be seen, some areas in the figure show that the simplified measurement of investors' opinion successfully captured some movements in terms of stock abnormal returns. However, unmatched areas appear in the figure at the same time. In this chapter, a forecasting framework is developed for stock price

shocks, which focuses on enhancing its ability of social attention modeling with sentiment analysis. The validated approach would facilitate diverse tasks for both policymakers and private market participants.

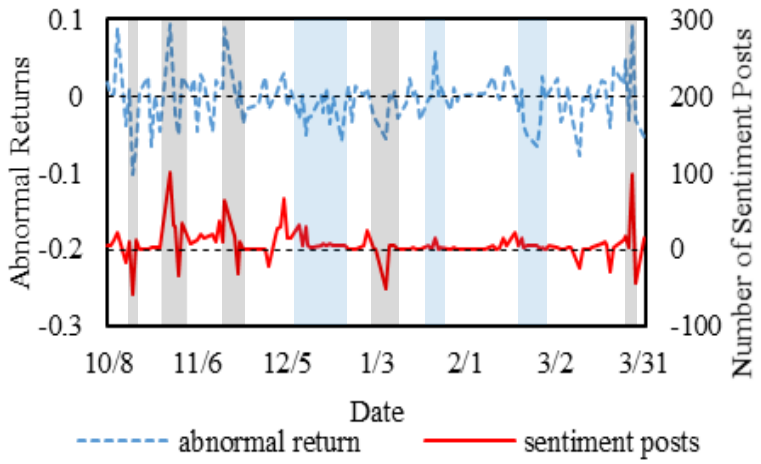


Figure 17. Abnormal Returns and Number of Opinion Posts for Stock 600643

5.1 Proposed Framework

In finance, a price shock, or stock abnormal return, can be defined by the difference between the expected return and its actual value. It is a portion of market movement that cannot be explained by financial models. I propose to generate a new social attention indicator to capture and predict these price shocks by leveraging on powerful classification techniques. The problem is formalized as follows.

As discussed before, given a dynamic social media network $\mathcal{G} = (\mathcal{N}, E)$, $\mathcal{I}$ is the set of historical information flows within the network for stock $q = 1, 2, 3, \dots Q$ at time $t = 1,2, \dots, T$. At time t, for a given stock, a list of features $X_{q,t} = \{x_1(q,t), x_2(q,t), \dots\}$ is formed based on $\mathcal{I}$ and other market information. The goal is to predict the price shock of stock q for the next period, notated by $Y(q; t+1)$, which is set to be three classes: Negative, Near-Zero, and Positive. The setting of the three categories is discussed in details below.

Figure 18 shows our proposed framework to solve the formalized problem. The framework is consisting of three major components: 1) Social Influence Modeling. Based on our earlier social influence propagation model [28, 41], we calculate social influence $g(i,j,t)$ and update self-influence $g(i,i,t)$ at time t. 2) Sentiment Analysis. A sentiment factor is coupled with our earlier enhanced DSA framework to get more precisely values by identifying the positive and negative keywords in each article accordingly. Combined with $g(i,j,t)$, we get the sentiment-DSA, notated by $a_q(t)$. 3) Price Shock Forecasting. With $a_q(t)$ as an import feature dimension plus other historical information, we apply different classifiers to predict the future price shocks' categories $Y(q; t+1)$. In additional to the sentiment-DSA, our input features also include information come from the stock market: historical price, historical returns, historical trading volume, historical beta risks, historical abnormal returns, et al.

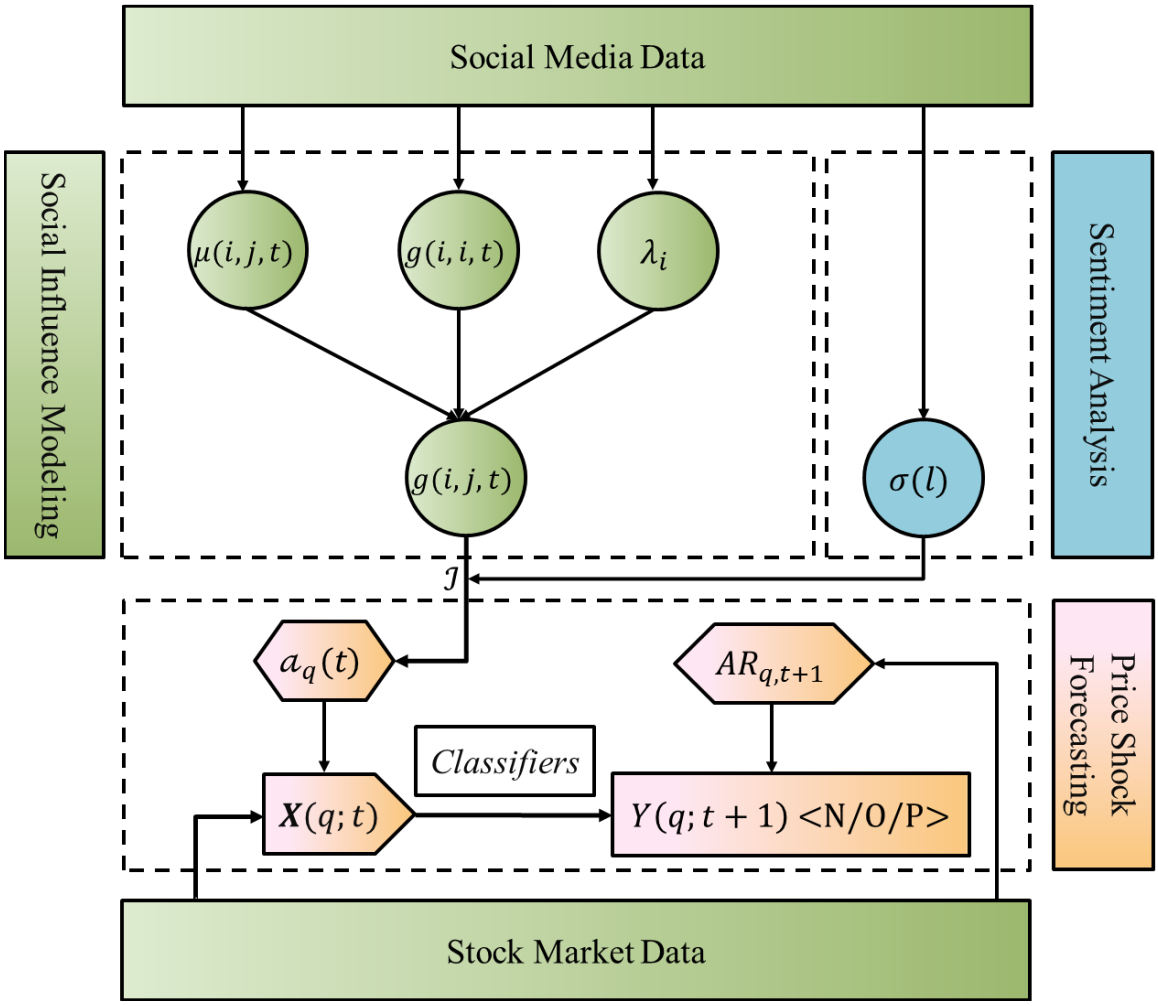


Figure 18. Proposed Framework of the DSA-based Classification Approach

This framework is based on analyses before that launched deep studies of financial activities in social media and their essential impacts on the stock market, and introduced the degree of social attention (DSA). The effectiveness of DSA in capturing stock price shocks was verified.

Limitations, however, still exist in practice. First, sentiment information was not considered in this measurement, hence it could not tell the actual direction of the prices shock. Second, the forecasting power of DSA was not checked.

5.2 Sentiment Analysis

In chapter 2, the DSA is defined as **Equation (10)**. In addition to the original DSA, the sentiment indicator is taken into consideration, so that the new measurement – Sentiment DSA - would be able to learn from the positive and negative keywords in each article accordingly, and actually capture its opinion [49].

In order to include the public opinion of one article posted by node i at time t, a sentiment factor $\sigma(l)$ for each article l among $d_q(i;t)$ is added and **Equation (10)** to yield a new DSA measurement as:

$$a'_q(t) = \sum_{i=1}^{N} \sum_{l=1}^{d_q(i;t)} \sigma(l) \sum_{j=1}^{N} \varphi(i,j;t) f_q(i,j;t) \tag{31}$$

This new measurement considers the opinion of each social media article, and it is used as the key variable in the forecasting framework. In the equation, the value of $\sigma(l)$ is the sentiment score of an article l.

It is relied on a widely used sentiment dictionary [48] that differentiates the positive and negative keywords/phrases and provides sentiment score for each one of them. The sentiment score for each post is calculated as:

$$\sigma(l) = \Phi(l) + \Psi(l) \tag{32}$$

where $\Phi(l)$ is the total sentiment score of positive keywords in article l and $\Psi(l)$ is the total sentiment score of negative ones. Especially, the used sentiment dictionary considers multiple combination of keywords and provides different sentiment scores to them.

There are some important stuffs needed to be noticed when applying our method to every articles. First, one article may discuss multiple stocks and sentiments for each one might be different. However, with our stock-based approach, we only select articles related to one specific stock in our DSA framework. So the effects of this problems are reduced and it is more relevant in our study.

Second, simple keyword counting can easily make mistakes. In the dictionary, it considers multiple combination of keywords and assign powers to each term. In this way, the chance of making mistakes in sentiment is eliminated to a certain degree and the calculation is made more precisely.

5.3 Classification Results

In this section, the model is evaluated with two aspects: 1) compare the performance of classification of features with the DSA terms and without DSA terms to show the improvement of our method; 2) in addition, list several major classifier algorithms to present the effectiveness of our approach.

5.3.1 Experiment Setups

This study consists of three main components: 1) perform a feature selection process to identify the suitable feature terms for stock abnormal return classification; 2) construction of a classifier for predicting daily abnormal returns; 3) evaluation of the performance of the classifier by comparing different classifier methods and feature terms with our DSA framework.

The first step is to determine the positive and negative abnormal returns among our data. **Figure 19** shows an example about the histogram plot of abnormal returns for a Chinese stock (601106) in stock market. As we know, the abnormal returns for one stock can be simply assumed to have the normal distribution so the abnormal returns can be divide into three classes, Negative (N)/Near-Zero (O)/Positive (P). For the abnormal returns below minus one standard deviation can be considered as class “Negative (N)” and for the abnormal returns above one standard deviation of full sample can be considered as class “Positive (P)”. The abnormal returns appear within one standard deviation can be considered as class “Near-Zero (O)” accordingly.

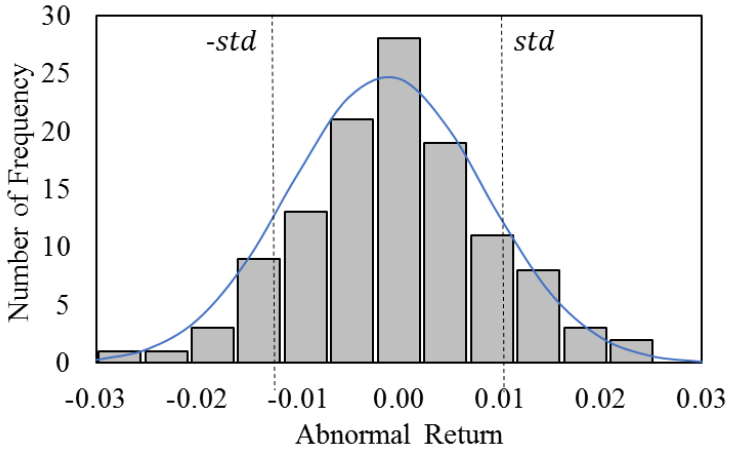


Figure 19. An Example: Histogram of Abnormal Returns for Stock (601106)

Several financial features from stock market dataset are examined and I conduct a feature selection process provided by a software platform called Weka [50], a collection of machine learning algorithms for data mining tasks. Based on the feature selection results are consistent with the experimental results [28, 41], the historical returns and trading volumes are selected as the most important variables in the learning process for the abnormal return classification. According to the discussions of time effects, it has 5 weighted DSA terms as the input based on the results in [41].

In a classification task, the Precision for a class is the number of true positives divided by the total number of elements labeled as belonging to the positive class. Recall in this context is defined as the number of true positives divided by the total number of elements that actually

belong to the positive class. The two measures are used together in the F-measure to provide a single measurement for a system [51]. The F-measure is defined as follow:

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{33}$$

Especially in this work, the elements in class "O" are taken large fractions of the whole population and we are most interested in the identification of class "N" and "P". So additionally the F-measure of class "Near-Zero (O)" excluded is formed for a better illustration of the performance. Followed by the classification procedure, Naive Bayes [52], Decision Tree (J48) [53], Radom Forest [54, 55], Logistic [56, 57] and LibSVM [58, 59] are the chosen classifier algorithms to train the data. The 10-fold cross-validation is used in the evaluation and the experiments are performed in Weka accordingly [60].

5.3.2 Comparison of Classifiers

The positive relationship between DSAs and stork abnormal returns in Chinese stock market has been proved in [28, 41]. Now, we test how the DSAs would help in stock abnormal returns prediction.

In **Table 15**, Panel A reports results of the total F-measures and Panel B reports the F-measures with results of class "O" excluded. As can be seen, Naive Bayes, Decision Tree (J48) and Radom Forest perform better than the other two methods, Logistic and LibSVM. For the three top

classifiers in **Table 15**, all the results are within good values about 0.80 in Panel A and 0.70 in Panel B, and it gives us strong evidence that our DSA framework can be used as an important factor to identify the stock abnormal returns. Among them, the smallest F-measures are in the "service" subsample (F-measure = 0.801) in Panel A and the "utility" subsample (F-measure = 0.681) in Panel B.

Among all the three better classifiers, it can be seen that the results from Random Forest is mostly better than the other two in both Panel A and B. For Naïve Bayes, the average F-measures is about 0.807 in Panel A and the minimum value is found from the subsample of "S" equals to 0.801. The average F-measures is about 0.689 in Panel B and the minimum value is found from the subsample of "U" equals to 0.681. For Decision Tree (J48), the average F-measures is about 0.813 in Panel A and the minimum value is found from the subsample of "BM" equals to 0.806. The average F-measures is about 0.699 in Panel B and the minimum value is found from the subsample of "F" equals to 0.693. For Random Forest, the average F-measures is about 0.820 in Panel A and the minimum value is found from the subsample of "CG" equals to 0.814. The average F-measures is about "0.715" in Panel B and the minimum value is found from the subsample of "S" equals to 0.700.

Table 15. F-measures for Different Classifiers

Panel A: All classes included							
	BM	CG	F	IG	S	U	Total
Naïve Bayes	0.803	0.802	0.815	0.814	0.801	0.807	0.807
Decision Tree (J48)	0.806	0.809	0.819	0.818	0.813	0.811	0.813
Random Forest	0.824	0.814	0.822	0.819	0.819	0.819	0.820
Logistic	0.781	0.79	0.765	0.765	0.772	0.78	0.776
LibSVM	0.559	0.601	0.574	0.583	0.552	0.599	0.578

Panel B: Class “Near-Zero (O)” excluded							
	BM	CG	F	IG	S	U	Total
Naïve Bayes	0.688	0.703	0.683	0.692	0.69	0.681	0.689
Decision Tree (J48)	0.708	0.705	0.693	0.700	0.694	0.694	0.699
Random Forest	0.712	0.719	0.720	0.723	0.701	0.715	0.715
Logistic	0.651	0.666	0.661	0.659	0.653	0.657	0.658
LibSVM	0.283	0.311	0.272	0.29	0.3	0.301	0.293

Another interesting finding showed in **Figure 20** is that negative attention would better predict stock price shocks than positive one. As can be seen, based on the total sample results, the results of the top three performed classifiers including Naïve Bayes, Decision Tree (J48), and Random Forest are showed. This evidence is consistent with the findings in several previous works that conclude that negative words in an article captures the tone of the report.

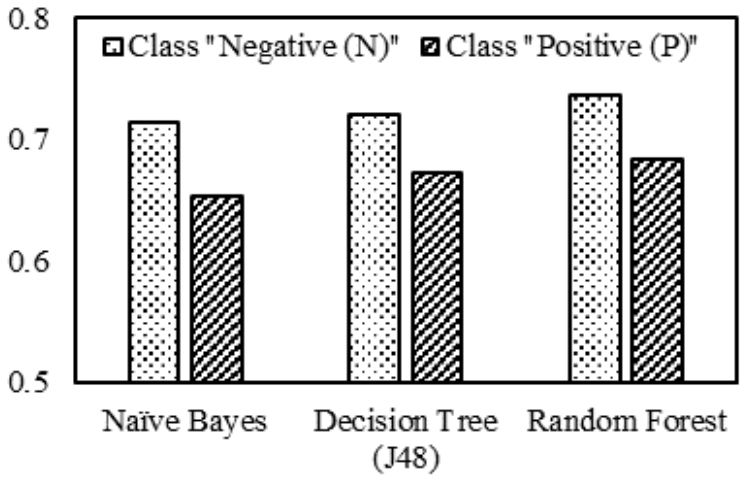


Figure 20. Comparison of the forecasting performance for class “Negative (N)” and class “Positive (P)” with total sample

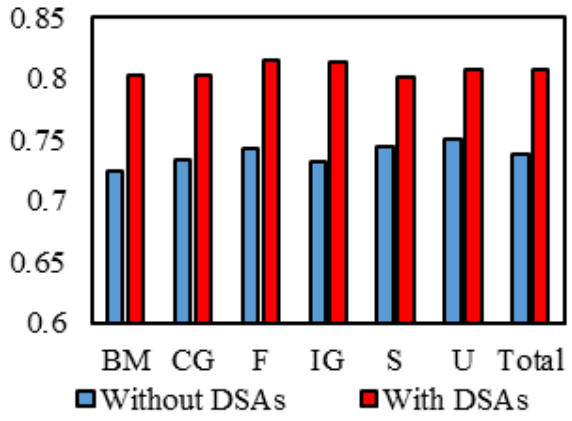
5.3.3 Testing the Impact of DSA

In this section, the improvements of DSAs on stock abnormal returns prediction is studied by comparing classification performance results of feature terms with and without sentiment DSA terms respectively. In addition to the test based on full sample, the data are separated into several

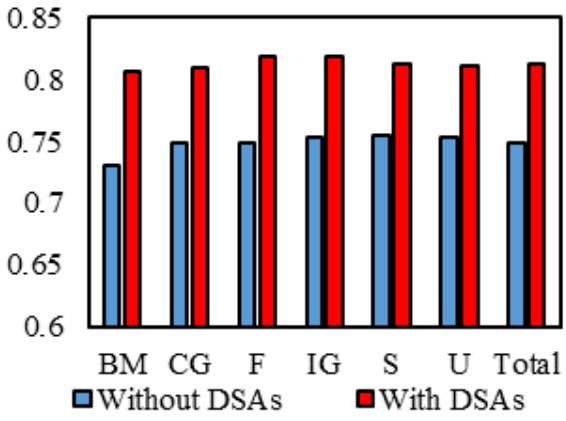
groups based on business sectors. The same three classifiers are selected, Naive Bayes, Decision Tree (J48) and Radom Forest for their better performance, as discussed in the last section.

Figure 21 and **Figure 22** plots the estimated F-measures of the three classifier methods based on seven samples. Additionally, Figure shows the F-measures with results of class "O" excluded as discussed before. **Figure 21(a)** and **Figure 22(a)** report the test results come from the Naïve Bayes, **Figure 21(b)** and **Figure 22(b)** report the results come from the Decision Tree (J48) and **Figure 21(c)** and **Figure 22(c)** report the results come from the Random Forest. Several empirical findings can be concluded from the results. Additionally, **Figure 21(d)** and **Figure 22(d)** report the results with error bar of all these classifiers for the comparison of total sample.

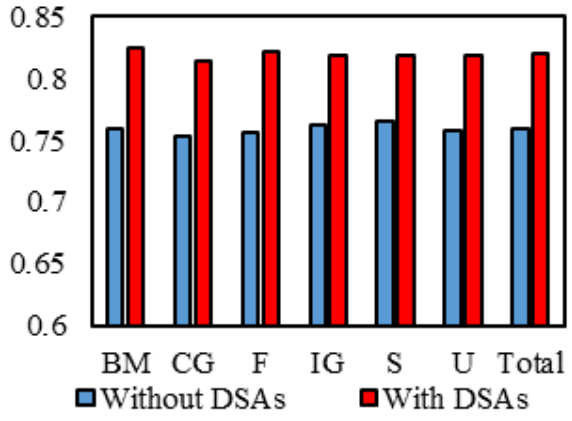
First, I report the F-measures for each sample and see significant improvements in terms of the performance of data fitting with DSA terms. More importantly, I find that the F-measures with DSA terms is greater than the ones without DSA terms. It can be seen that all F-measures without DSA terms in **Figure 21** are about 0.75 and all F-measures with DSA terms are about 0.81. For the F-measures with results of class "O" excluded, results without DSA terms are about 0.52 and results with DSA terms are about 0.70 in **Figure 22**. These results support that our DSAs improve the abnormal returns predictions accordingly.



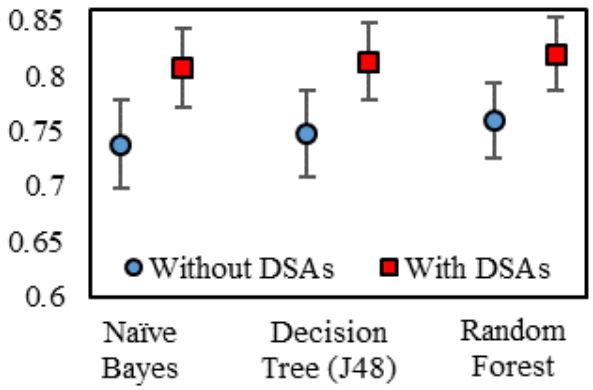
(a) Naïve Bayes



(b) Decision Tree (J48)

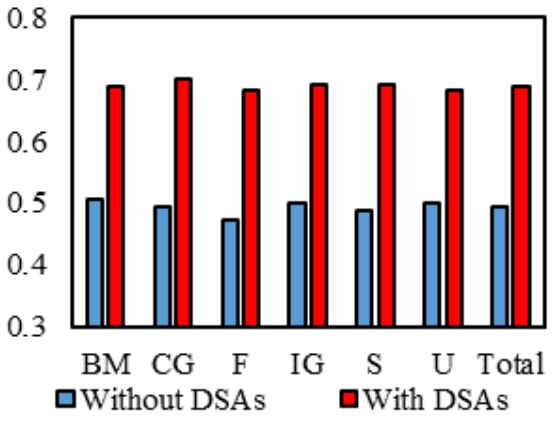


(c) Random Forest

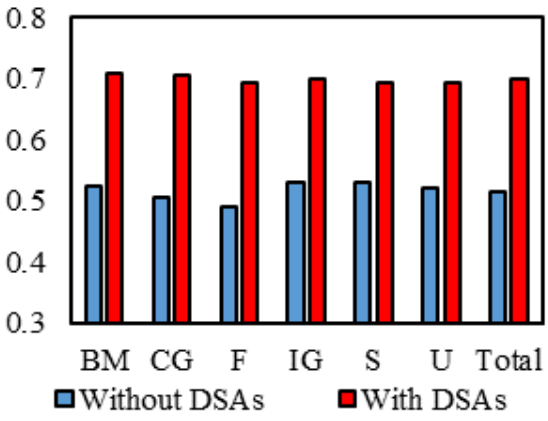


(d) Comparison of Total Sample

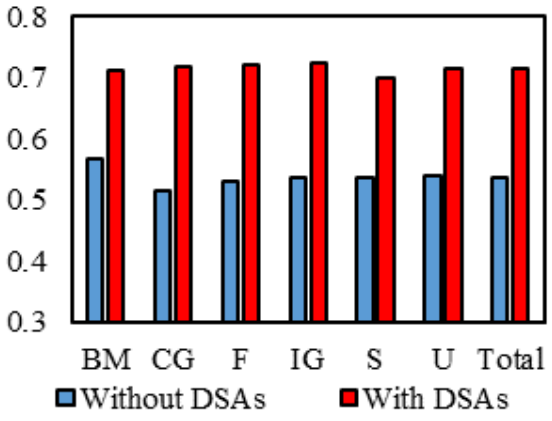
Figure 21. F-measure of all classes included for the impact of DSA



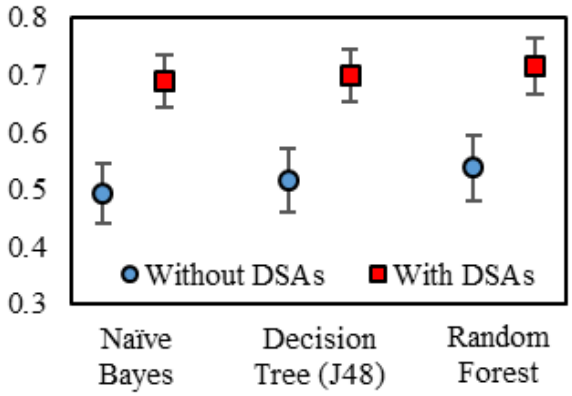
(a) Naïve Bayes



(b) Decision Tree (J48)



(c) Random Forest



(d) Comparison of Total Sample

Figure 22. F-measure of class “Near-Zero (O)” excluded for the impact of DSA

Second, I find that the F-measures with results of class “O” excluded in Figure can clearly show the predict power of DSA terms for certain class as expected. As shown in the **Figure 21**, comparing with results without DSA terms, it can be seen that a significant improvement in results with DSA terms as well as in the **Figure 22**. Therefore, it believes that our sentiment DSA framework effectively enhance the prediction of stock price shocks in certain classes.

Chapter 6 Conclusions

In this research work, I propose a Degree of Social Attention (DSA) framework for stock analysis. The goal is to identify the essential linkage between social media activities and the stock market. To do, a measurement of stock social attention is proposed by extend the influence model introduced by [27]. The original model is improved by (1) mathematically defining the self-influence (confidence) and (2) designing algorithms to capture dynamic influence in a social network with the consideration of computational efficiency. The DSA based on the proposed dynamic influence process is further defined.

By testing our two hypotheses formulized under the assumptions of semi-strong inefficiency of the Chinese stock market, the effectiveness of the DSA framework is verified with both the market-based approach and the stock-based approach. Positive relationship is found between the absolute value of abnormal return and DSA. It also proves that the market volume is associated with DSA at most of time. On the other hand, the results can serve as new evidence to support the semi-strong market inefficiency in the Chinese stock market, and also show the significance of DSA as an important factor to link social media activities to the stock market.

Furthermore, an influence propagation model based on Gamma distribution is proposed to estimate the time effect of previous social attention, and show that our data is well fitted by the model. I investigate the effects of historical DSAs on the stock abnormal returns based on the weighted DSAs. The results serve as additional evidence to support the significance of current

DSA in terms of their influence on the stock abnormal returns. And evidence is found to show that historical DSAs significantly affect stock abnormal returns. In addition, to investigate the fitting performance, the adjusted R-squares are checked and compared for all the models. The results show that the models that take the historical DSAs into account significantly outperform the ones that solely consider the current DSA. The results show that considering historical DSAs significantly improves the ranking performance in price shocks detection as well. Consistent results are also found in the comparison of the adjusted R-squared. The models that take the historical DSAs into account significantly outperform those that solely consider the current DSA.

Additionally, I propose a sentiment DSA measurement that considers opinions from social media authors as well as their social influence in a social network. The effectiveness of the measurement is evaluated by investigating the performance of using it to handle stock price shocks forecasting tasks, which formalized as a classification problem. Based on the results, several findings are concluded as follows. First, including the sentiment DSA in the framework would consistently improve the performance of forecasting stock price shocks. Second, within the five tested classifiers, Random Forest algorithm perform the best under all samples, and SVM has the worst performance. Third, comparing to positive social attention, the negative one would result better forecasting results.

Bibliography

[1] E. F. Fama, "Efficient capital markets: A review of theory and empirical work*," *The journal of Finance,* vol. 25, pp. 383-417, 1970.

[2] E. F. Fama and K. R. French, "Dissecting anomalies," *The Journal of Finance,* vol. 63, pp. 1653-1678, 2008.

[3] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of Finance,* vol. 48, pp. 65-91, 1993.

[4] N. Jegadeesh and S. Titman, "Profitability of momentum strategies: An evaluation of alternative explanations," *The Journal of Finance,* vol. 56, pp. 699-720, 2001.

[5] A. Agrawal, J. F. Jaffe, and G. N. Mandelker, "The post-merger performance of acquiring firms: a re-examination of an anomaly," *The Journal of Finance,* vol. 47, pp. 1605-1621, 1992.

[6] R. Ball and P. Brown, "An empirical evaluation of accounting income numbers," *Journal of accounting research,* pp. 159-178, 1968.

[7] P. L. Davies and M. Canes, "Stock prices and the publication of second-hand information," *Journal of Business,* pp. 43-56, 1978.

[8] A. Brav, C. Geczy, and P. A. Gompers, "Is the abnormal return following equity issuances anomalous?," *Journal of Financial Economics*, vol. 56, pp. 209-249, 2000.

[9] T. O. Sprenger, A. Tumasjan, P. G. Sandner, and I. M. Welpe, "Tweets and Trades: the Information Content of Stock Microblogs," *European Financial Management*, vol. 20, pp. 926-957, 2014.

[10] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of internet stock message boards," *Journal of Finance*, vol. 59, pp. 1259-94, 2004.

[11] S. Das, A. Martinez-Jerez, and P. Tufano, "eInformation: a clinical study of investor discussion and sentiment," *Financial Management*, vol. 34, pp. 103-37, 2005.

[12] S. Sabherwal, S. K. Sarkar, and Y. Zhang, "Online talk: does it matter?," *Managerial Finance*, vol. 34, pp. 423-36, 2008.

[13] J. L. Koski, E. M. Rice, and A. Tarhouni, "Noise trading and volatility: Evidence from day trading and message boards," 2004.

[14] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *the International Conference on Weblogs and Social Media*, 2010, pp. 10-17.

[15] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: conversational aspects of retweeting on Twitter," in *the 43rd Hawaii International Conference on System Sciences*, 2010, pp. 1-10.

[16] B. Gu, P. Konana, and H.-W. Chen, "Melting-pot or homophily? An empirical investigation of user interactions in virtual investment-related communities," 2008.

[17] J. Cai, R. A. Walkling, and K. Yang, "The price of Street friends: Social networks, informed trading, and shareholder costs," *Journal of Financial and Quantitative Analysis (JFQA), Forthcoming*, 2014.

[18] O. Faleye, T. Kovacs, and A. Venkateswaran, "Do Better-Connected CEOs Innovate More?," *Journal of Financial and Quantitative Analysis*, vol. 49, pp. 1201-1225, 2014.

[19] H. Chen, P. De, Y. J. Hu, and B.-H. Hwang, "Wisdom of crowds: The value of stock opinions transmitted through social media," *Review of Financial Studies*, vol. 27, pp. 1367-1403, 2014.

[20] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, vol. 12, pp. 211-223, Aug 2001.

[21] M. Granovetter, "Threshold Models of Collective Behavior," *American Journal of Sociology*, vol. 83, pp. 1420-1443, 1978.

[22] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137-146.

[23] B. Hayo and A. M. Kutan, "The impact of news, oil prices, and global market developments on russian financial markets.," *The Economics of Transition*, vol. 13, pp. 373-393, 2005.

[24] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, "Mining of concurrent text and time series.," in *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, 2000, pp. 37-44.

[25] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system. ," *ACM Transaction Information Systems*, vol. 27:12:1-12:19, 2009.

[26] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data.," *Nature*, vol. 457, pp. 1012-1014, 2009.

[27] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang, "Pagerank with priors: An influence propagation perspective," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013, pp. 2740-2746.

[28] L. Zhang, K. Xiao, Q. Liu, Y. Tao, and Y. Deng, "Modeling Social Attention for Stock Analysis: An Influence Propagation Perspective," in *Proceedings of the 15th IEEE International Conference on Data Mining*, Atlantic City, NJ, 2015.

[29] M. Sosonkina, L. T. Watson, and R. K. Kapania, "A new adaptive GMRES algorithm for achieving high accuracy," 1996.

[30] A. S. Kyle, "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society*, pp. 1315-1335, 1985.

[31] M. C. Jensen, "Problems in Selection of Security Portfolios - Performance of Mutual Funds in Period 1945-1964 - .1. Introduction," *Journal of Finance*, vol. 23, pp. 389-416, 1968.

[32] C. R. Rao, "Citation Classic - Linear Statistical-Inference and Its Applications," *Current Contents/Social & Behavioral Sciences*, pp. 14-14, 1980.

[33] F. Hayashi, *Econometrics*. Princeton: Princeton University Press, 2000.

[34] E. F. Fama and K. R. French, "The capital asset pricing model: Theory and evidence," *Journal of Economic Perspectives*, vol. 18, pp. 25-46, Sum 2004.

[35] W. F. Sharpe, "Citation Classic - Capital Asset Prices - Theory of Market Equilibrium under Conditions of Risk," *Current Contents/Social & Behavioral Sciences*, pp. 12-12, 1979.

[36] S. Ma, *The efficiency of China's stock market*: Ashgate Aldershot, 2004.

[37] HtmlUnit. *HtmlUnit* *http://htmlunit.sourceforge.net/*.

[38] R. P. Fishe and A. D. Smith, "Identifying informed traders in futures markets," *Journal of Financial Markets,* vol. 15, pp. 329-359, 2012.

[39] C. M. Corcoran, *Long/short market dynamics: trading strategies for today's markets* vol. 323: John Wiley & Sons, 2007.

[40] CSI300, "Shanghai Shenzhen CSI 300 http://en.wikipedia.org/wiki/CSI_300_Index."

[41] L. Zhang, K. Xiao, Q. Liu, and C. Liu, "Influence-Based Social Attention Modeling for Price Shocks Detection," *ACM Transactions on Management Information Systems (TMIS),* 2016.

[42] L. Zhang, L. Zhang, X. Keli, and Q. Liu, "Forecasting Price Shocks with Social Attention Modeling and Sentiment Analysis," presented at the The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016), San Francisco, CA, USA, 2016.

[43] N. C. Avery, A. J. Chevalier, and J. R. Zeckhauser, "The "CAPS" Prediction System and Stock Market Returns," *Review of Finance,* vol. 1, 2015.

[44] S. K. Deng, T. Mitsubuchi, and A. Sakurai, "Stock Price Change Rate Prediction by Utilizing Social Network Activities," *Scientific World Journal,* 2014.

[45] T. H. Nguyen and K. Shirai, "Aspect-Based Sentiment Analysis Using Tree Kernel Based Relation Extraction," *Computational Linguistics and Intelligent Text Processing (Cicling 2015), Pt Ii,* vol. 9042, pp. 114-125, 2015.

[46] G. Wang, T. Y. Wang, B. L. Wang, D. Sambasivan, Z. B. Zhang, H. T. Zheng, *et al.*, "Crowds on Wall Street: Extracting Value from Collaborative Investing Platforms," *Proceedings of the 2015 Acm International Conference on Computer-Supported Cooperative Work and Social Computing (Cscw'15),* pp. 17-30, 2015.

[47] D. Simon and R. Heimer, "Facebook finance: How social interaction propagates active investing," presented at the AFA 2013 San Diego Meetings, San Diego, CA, 2012.

[48] L. Xu, H. Lin, and Y. Pan, "Constructing the Affective Lexicon Ontology," *Journal of The China Society For Scientific and Technical Information,* vol. 27, pp. 180-185, 2008.

[49] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications,* vol. 42, pp. 9603-9611, Dec 30 2015.

[50] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, 2009.

[51] E. C. Jensen, S. M. Beitzel, A. Chowdhury, and O. Frieder, "Query phrase suggestion from topically tagged session logs," *Flexible Query Answering Systems, Proceedings*, vol. 4027, pp. 185-196, 2006.

[52] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (3rd ed.)*: Prentice Hall, 2009.

[53] J. R. Quinlan, "Simplifying Decision Trees," *International Journal of Man-Machine Studies*, vol. 27, pp. 221-234, Sep 1987.

[54] T. K. Ho, "Random Decision Forests," in *the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 1995, pp. 278–282.

[55] T. K. Ho, "The random subspace method for constructing decision forests," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832-844, Aug 1998.

[56] D. A. Freedman, *Statistical Models: Theory and Practice*: Cambridge University Press, 2009.

[57] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression (2nd ed.)*: Wiley, 2000.

[58] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443-1471, Jul 2001.

[59] B. Scholkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation,* vol. 12, pp. 1207-1245, May 2000.

[60] P. A. Devijver, "Standardization - Mathematical-Methods in Assortment Determination - Bongers,C," *European Journal of Operational Research,* vol. 9, pp. 310-311, 1982.