Expert Systems with Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A social-media-based approach to predicting stock comovement

Ling Liu [a], Jing Wu [a], Ping Li [a], Qing Li [a,b,*]

[a] School of Economics Information Engineering, Southwestern University of Finance and Economics, China
[b] Management Information Systems Department, University of Arizona, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Social media
Comovement
Microblogging
Industry classification

ABSTRACT

Stock return comovement analysis is important to financial analysts, decision makers, and academic researchers and has many financial implications, such as portfolio management, style investing, and market risk detecting. This paper proposes a novel model to both identify homogeneous stock groups and predict stock comovement with respect to firm-specific social media metrics. One of the innovations of the social media platform is that it breaks traditional media intermediation. A firm with an official Twitter account can publish information and interact with its users directly. Such direct information is largely reflected on firm-specific metrics, e.g., the firm's number of followers and number of tweets sent. To the best of our knowledge, this paper is the first to reveal the impact of social media metrics on stock return comovement studies. By analyzing samples from the NYSE and NASDAQ stock exchanges, we find that firms with official Twitter accounts have a much higher comovement than those without such accounts. Furthermore, we classify the former set of firms into homogeneous groups by their specific microblogging metrics. The results demonstrate that these metrics cannot only predict the comovement of stocks but also notably increase the accuracy of comovement predicting, compared with industry categories.

1. Introduction

Stock return comovement analysis refers to identifying homogeneous groups of stocks that have similar movement of returns. Such an analysis is very important to financial analysts, decision makers, and academic researchers. On a practical level, the analysis of homogeneous stock groups can help investors construct investment portfolios. To minimize risks, many investors allocate their funds across different homogeneous assets groups (Barberis & Shleifer, 2003; Chan, Lakonishok, & Swaminathan, 2007; Swensen, 2009). Stock return comovement is also inversely associated with economic resources allocation (Chelley-Steeley, Lambertides, & Savva, 2013; Durnev, Morck, & Yeung, 2004; Wurgler, 2000).

Although many factors lead to similar stock returns, such as firm size, industry, investor behavior, and market characteristics (Aghabozorgi & Teh, 2014; Brockman, Liebenberg, & Schutte, 2010; Chelley-Steeley et al., 2013; Claessens & Yafeh, 2012), we believe that it is still necessary to propose a new model to identify homogeneous stock groups from a social media perspective. Social media has changed how society communicates and organizes itself (Aral, Dellarocas, & Godes, 2013; Golbeck, Grimes, & Rogers, 2010; Yu & Kak, 2012). In particular, microblogging service platforms allow users to follow each other and, thus, to bypass the intermediate channel of information distribution between firms and the public. Communication among the public is also accelerated and enhanced by microblogging services. It is reasonable to infer that the stock market is not immune to the impact of such activities. In fact, a great deal of research has shown that social media users' activities can be used to predict firms' financial performances (Bollen, Mao, & Zeng, 2011; Yu, Duan, & Cao, 2013). We further suggest that such public behavior toward firms can be used as a proxy to predict the comovement between stocks.

By identifying official Twitter accounts for each publicly traded firm, our empirical results demonstrate that firms with official Twitter accounts have much higher comovement than firms without such accounts. This finding is similar to the results of Barberis, Shleifer, and Wurgler (2005), which showed that when added to a major index, firms experience higher comovement because investors have better information access to the firms after they were added to the index. This might indicate that Twitter can enhance the diffusion of firm information, resulting in the same effects those caused by major stock indexes.

Furthermore, this paper proposes a novel model to identify homogeneous stock groups and to predict stock comovement with respect to firm-specific microblogging metrics. Unlike most

* Corresponding author at: Management Information Systems Department, University of Arizona, USA.
E-mail address: liq_t@swufe.edu.cn (Q. Li).

http://dx.doi.org/10.1016/j.eswa.2014.12.049

previous studies (see Bollen et al., 2011; Luo, Zhang, & Duan, 2013; Yu et al., 2013), which focus on firms' overall social media metrics, we concentrate on firm-specific metrics. Overall metrics refers to a firm's broad metrics on the entire platform, e.g., how many tweets have mentioned the firm's name, or the public mood associated with all firm-related tweets. Our research identifies each listed firm's microblogging account and examines their specific metrics. This approach has several advantages. First, retrieving overall metrics requires creating filters that use stock symbol, company name, and other keywords to identify firm-related information (Ruiz, Hristidis, Castillo, Gionis, & Jaimes, 2012). Such identifiers and criteria can be ambiguous, and a great deal of noise information might be collected, which in turn should be carefully handled. Firm-specific metrics, however, are entirely relevant to each firm. Second, retrieving overall metrics requires managing a large volume of data. Thus, most studies can only choose companies in limited industries to study, which may impact the generalization of results (Luo et al., 2013). Our approach enables us to analyze a larger dataset of firms from all of the industries listed on NYSE and NASDAQ stock exchanges.

Based on the dataset, we take key firms' microblogging metrics, e.g., the number of followers, as observations; and we then cluster those firms into homogeneous groups. The idea of clustering stocks into groups is already well accepted. For example, Farrell (1975) heuristically partitions stocks into four homogeneous groups and discusses their implications on portfolio construction, performance measurement, and group rotation. Aghabozorgi and Teh (2014) propose a three-phase clustering method to automatically categorize stock markets into different groups. Liao and Chou (2013) adopt a data mining approach to investigating the comovement between Taiwan and Hong Kong stock markets. By applying the model, our results show that firms in the same clustered groups have higher comovements than those in the same industry categories. Our research proves that with simple metrics, social media can produce better results compared to other traditional industry classifications. Another advantage of our model is that our proposed metrics are easy to be understood and verify, and they might attract more practical usage than complex models.

The remainder of the paper is structured as follows. Section 2 introduces the background and theory of the paper, and is followed by Section 3, which describes the model. Section 4 describes the data collection and research framework. Section 5 discusses the results. Finally, Section 6 summarizes the paper.

2. Theory and related work

2.1. Stock comovement

Comovement, as a component of market beta in the form of covariance, exerts an important influence over portfolio performance, the market risk premium, and the cost of capital (Chelley-Steeley et al., 2013; Roll, 1988). Thus, it has attracted a great deal of attention in many areas and is surrounded by a long-standing debate on its causes. Efficient Market Hypothesis (EMH) theory suggests that a firm's stock price reflects all the information and is thus only influenced by the firm's fundamental values (Fama, 1970). It is grounded on two basic assumptions: (1) all investors are rationals, and (2) stock markets are informationally efficient. According to this theory, the prices of different stocks move together only in response to common movements in their fundamental values.

However, these two assumptions have been consistently challenged by behavioral finance research. For instance, Rashes (2001) has discovered that even two assets with nothing in common but their similar names have abnormal comovement, caused by investors who have misidentified the assets' names. This research indicates that investors are not always rational and that their behavior may cause stock comovement. Pindyck and Rotemberg (1993) have found that company size and degree of institutional ownership also influence stock comovements. The work of Dutt and Mihov (2013) has shown that markets with similar industries demonstrate higher stock comovements. Such effects occur because many investors prefer to group stocks into categories based on their basic characteristics, such as firm size and industry categories, and then allocate funds at the level of those categories (Barberis et al., 2005). Interactions among investors are also found to have an impact on stock comovement. Based on Internet message discussion boards, Antweiler and Frank (2004) have demonstrated that investors' activities influence trading volume and return volatility. Chue, Gul, and Mian (2014) have find that aggregated stock market comovement becomes higher when investor sentiment is extreme.

Uneven information diffusion is another major factor that is considered to have an impact on stock return comovement. The media, the major information diffusion intermediate, plays an important role in stock price fluctuation (Li et al., 2014a). Quantified media content is proven to be able to predict stock prices (Chan, 2003; Tetlock, 2007). Researchers also have found that uneven diffusion of information can inspire stock comovement. When added to a major index, stocks experience significant increase in their betas (Vijh, 1994). The reason for such a change is that investors may have better access to information related to stocks in those major indexes (Claessens & Yafeh, 2012).

A great deal of research is based on media analysis which was once the only channel for information distribution and acquisition. Furthermore, prior to the social media era, it was difficult to collect investors' firm-related activities toward, which makes media a good anchor. The emergence of social media changed this situation.

2.2. Social media and firm financial performance

Social media has dramatically changed how people obtain information. In particular, microblogging services, which allow users to follow each other, to make comments and to forward tweets, provide a more flattened and efficient way for people to acquire and distribute information. Moreover, microblogging platforms also enable us to directly collect the public's opinions and behavior toward firms. Although tweeting may not provide new firm information to the market, firm tweets are found to increase information dissemination (Alexander & Gentry, 2014). Recent research has demonstrated that users' social-media activities can be used to predict firms' financial outcomes (Goh, Heng, & Lin, 2013; Tirunillai & Tellis, 2012; Zheludev, Smith, & Aste, 2014). Based on joint mentions in Twitter's public tweets, Sprenger and Welpe (2011) suggest that news-based relatedness can help delineate meaningful industry groups.

The stream of research analyzing the relationship between public activities on social media platforms and firms' financial performance is very popular. Researchers from marketing, information systems, economics, and many other domains have explored the effectiveness of social media information. A great deal of research has found that consumers' social media activities and interactions can affect their consumption intentions, which in turn will influence firms' sales (Dellarocas, Awad, & Zhang, 2004; Lee, Shi, Cheung, Lim, & Sia, 2011; Phang, Zhang, & Sutanto, 2013; Zhu & Zhang, 2010). Bollen et al. (2011) find that the public's mood on Twitter can predict the Dow Jones Industrial Average. Luo et al. (2013) discovered that social media metrics are significant indicators of firm equity value. Sul, Dennies, and Yuan (2014) collect data from public Twitter posts about S&P 500 firms and assert that the positive or negative emotion of tweets about a specific firm are significantly related to that firm's stock returns. Yu et al. (2013) find that overall social media metrics have a stronger relationship with firm stock performance than do conventional media. More interestingly, these discoveries indicated that the public's social-media activities, not those of the investors, can also be used to predict stock prices. That is because investors' emotions can be influenced by public sentiments, and thus, their investment decisions are also impacted (Li et al., 2014b).

Our research proposes a novel model to identify homogeneous stock groups based on firm-specific social media metrics. Table 1 compares recent literature related to this paper.

There are two main gaps in this stream of research. First, very little research has focused on identifying stock comovement from a social media perspective. Some pioneer researchers have begun to use advanced data-mining technologies in their stock comovement analysis. Liao and Chou (2013) apply data-mining technologies to analyze national stock markets. Aghabozorgi and Teh (2014) propose a three-phase clustering model to classify stocks into different groups based on their historical stock prices. These approaches usually use traditional financial metrics rather than social media metrics as indicators. For instance, Chue et al. (2014) find that aggregated stock market comovement becomes higher when investor sentiment is extreme.The investors' sentiment proxies were aggregated based on six proxies: the closed-end fund discount, NYSE share turnover, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium (Baker & Wurgler, 2006). Social media platforms, particularly microblogging platforms allow the public to retrieve and exchange information. They play a very important role in information dissemination (Alexander & Gentry, 2014). Thus, it is necessary to analyze their impact on stock comovement.

Second, one of the innovations of social media platforms is that they break traditional media intermediation. A firm with an official Twitter account can publish information and directly interact with its users. Such direct information is largely reflected by firm-specific metrics, e.g., its number of followers and the number of tweets it has sent. However, most prior studies that focus on connecting social media metrics with firm financial performance only concentrate on firms' overall social media metrics, i.e., a firm's broad metrics on the entire platform. For example, Sprenger and Welpe (2011), one of the few studies to examine the relationship between firms' social media metrics and their financial performance, use joint mentions of firms in all public tweets as the indicator that identifies stock groups. The unique value of firm-specific metrics might be reduced if they are mixed into the mass of overall metrics. Our research aims to fill the gaps by adopting firm-specific metrics as indicators to identify stock comovement. To the best of our knowledge, our paper is the first to take such an approach to investigating social media metrics and firm performance.

Furthermore, our model is built on quantified metrics. These metrics are easy to retrieve and be understood. Text metrics usually require text- and sentiment-mining techniques. Conversely, quantified metrics can reduce the complication of algorithms.

3. Models

Based on the previous discussion, it can be found that little attention has been paid to the relationship between social media data and stock comovement. This paper proposes a model of utilizing social media information to predict stock comovement. The following sections will discuss the possible microblogging metrics groups that can be used to predict stock return comovement.

3.1. Microblogging metrics

Thus far, most studies have concentrated on using social media metrics to predict firms' stock prices and returns, whose findings echo the behavioral financial perspective. We further suggest that social media metrics should be able to identify stock comovements. We expect that the number of followers, one of the most important microblogging metrics, can predict stock comovement.

On microblogging service platforms, firms are allowed to set up official accounts, which may be followed by other users. The number of followers refers to the number of users who are following the account and demonstrates how many users have access to a firm's announcements and published information. It can be seen as an indicator of firms' information-diffusion abilities. Based on information-diffusion comovement theory, this indicator implies that firms with similar numbers of followers may have higher comovement than firms with different numbers of followers. Given that the U.S. Securities and Exchange Commission (SEC) permits companies to use social media sites, including Facebook and Twitter, to communicate company announcements (Gallu, 2013) and that the spread of information on such platforms is much more efficient than traditional ways, it is reasonable to assume that users' activities on social media may influence firms' stock performance and thus influence stock comovements.

Another reason that we expect the number of followers to reflect stock comovement is that it is similar to the concept of "firm celebrity". Firm celebrity is considered as a firm's intangible asset (Rindova, Pollock, & Hayward, 2006). Both theoretical and

Table 1
Literature comparison.

	Predictor/indicator	Method/model	Financial outcome
Aghabozorgi and Teh (2014)	Stock prices	Three-phase clustering model	Stock comovement
Chue et al. (2014)	Aggregated investor sentiment	Regression	Stock comovement
Liao and Chou (2013)	Category stock indexes	Cluster analysis	Comovement between national market indexes
Sprenger and Welpe (2011)	Public social media metrics (joint mentions in public tweets)	Clustering (fraction analysis)	Stock comovement
Bollen et al. (2011)	Public tweets	Sentiment analysis	Daily DJIA value
Chen et al. (2014)	Public comments	Regression	Stock returns and earnings surprises
Jiang et al. (2014)	Discussion forum	Sentiment analysis	Stock return, volatility, and trade volume
Li et al. (2014b)	Financial news and discussion board	Sentiment analysis, SVM	Stock return
Luo et al. (2013)	Web blog, consumer rating, web traffic and Google search	VARX	Stock return and risk
Yu et al. (2013)	Blogs, forum, microblogs, and Google news	Sentiment analysis, panel analysis	Stock return and risk
Sul et al. (2014)	Public tweets	Regression stock	Stock return
This paper	Firm-specific social media metrics	K-means clustering, pairwise correlation	Stock comovement

empirical research has shown that firm celebrity has positive effects on firm financial outcomes (Pfarrer, Pollock, & Rindova, 2010; Rindova et al., 2006) and stock market performance (Liu, Li, Xu, & Zhang, 2014). Firm celebrity has two defining characteristics: the public's attraction and positive emotion (Rindova et al., 2006). A user's action of following a firm's account on a microblogging platform shows his/her interest in and fondness for the firm (Kaplan & Haenlein, 2011). Thus, the number of followers, which indicates the account;s popularity and attractiveness, can be regarded as a measurement of the level of firm celebrity. The fact that the most-followed individual accounts on Twitter are those of celebrities such as Katy Perry, Lady Gaga and Barack Obama can be taken as evidence to support this measurement. Similarly, prior work has shown that consumer and public attention, such as website traffic (Luo & Zhang, 2013) and consumer visit frequency (Rishika, Kumar, Janakiraman, & Bezawada, 2013), have a positive relationship with firm stock performance.

Based on the above analysis, it is reasonable to assume that the number of followers can be used to classify homogeneous stock groups. Therefore, our first assumption is that the number of followers of a firm's official account (P-FO) can be used to predict stock comovement.

P-FO is not only related to a firm's fame, but is also influenced by that firm's effort on the microblogging platform. In addition to being followed by the public, firms can post tweets and follow other users. Usually the more active a firm is, the more followers it attracts (Kaplan & Haenlein, 2011). To reduce the effects of the firm's own activities, we use two ratios to compare the effects of P-FO: the ratio of the number of followers to the number of followings, and the ratio of the number of followers to the number of tweets, The number of followings is the number of accounts that a firm follows, whereas the number of tweets is how many messages a firm has sent. The two ratios exhibit a firm's ability to attract followers with each tweet or following action. Thus, our second group of microblogging metrics includes the following information: the ratio of the number of followers to the number of followings (RFF), and the ratio of the number of followers to the number of tweets (RFT).

Another problem with the number of followers is that it can be manipulated. Fake accounts can be purchased and programmed to follow a specific account (Elder, 2013). To compare the effects of fake followers, we introduce another group of metrics: the number of firm-followers (F-FO),[1] which refers to the number of followers that are listed firms. Those firms are all authentic users, as we identified each one.

3.2. Control metrics

In addition to social media metrics, we should consider both fundamental and category factors that may influence stock comovements. Firm size and industry sector are the two factors that have a huge impact on stock comovements (Aghabozorgi & Teh, 2014; Brockman et al., 2010; Chelley-Steeley et al., 2013; Claessens & Yafeh, 2012). Firm size is an important factor in the financial analysis of a company (Kumar, Rajan, & Zingales, 1999). Chan et al. (2007) have found that within the same industry, larger firms have better return correlations than do smaller firms. Thus, we add firms' market capitalization (MC) as an observation in coordination with the microblogging metric groups.

Researchers have also found that stocks in the same industry have a higher return comovement than stocks across industries (Barberis et al., 2005; Chan et al., 2007; Chou, Ho, & Ko, 2012; Dutt & Mihov, 2013; Lessard, 1974; Moskowitz & Grinblatt, 1999). To account for industry categories, we classify stocks based on their industries, i.e., we use the industry classification scheme as the base and further classify the within-industry firms into narrower groups. There are four commonly adopted industry classification schemes used by financial and academic professionals. Standard Industrial Classification (SIC) codes have three main levels. Industry divisions are represented by Divisions A, B, etc. Two-digit SIC codes indicate Major Groups, and four-digit codes denote Industry Groups. The North American Industry Classification (NASIC), which provides a greater level of detail, replaced the SIC codes in 1997. Beginning with the four-digit SIC codes, Fama and French (1997) have reorganized firms into 48 industry groups. Although the Fama–French scheme has been widely adopted by academic researchers in many financial areas, the GICS scheme is widely recognized by financial professionals. This scheme was developed by Standard & Poor's and MSCI/Barra in response to the needs of the global financial community (Standard & Poor's, 2006). Bhojraj, Lee, and Oler (2003) compared the four classification schemes based on how well they explain stock return comovements, cross-sectional variations, and various key financial ratios. The results show that GICS classifications are "significantly better at explaining stock return comovements". Similar results were obtained by Chan et al. (2007), who found that the six-digit GICS code has the strongest impact. Thus, we chose the GICS industry scheme as the industry classification, which groups stocks into 10 two-digit GICS sectors, 24 four-digit GICS industry groups, 67 six-digit GICS industries and 147 eight-digit GICS sub-industries. Considering that dividing the stocks into sub-industries, would result in a very small number of stocks two- and four-digit groups are selected as base groups.

In summary, we propose the following five metrics groups for predicting stock comovement by classifying stocks based on their two- and four-digit GICS industry categories. Our assumption is that the following five groups of metrics can predict stock comovements.

Group 1: P-FO, MC
Group 2: RFF, RFT, MC
Group 3: F-FO, MC
Group 4: F-FO, FO, MC
Group 5: F-FO, RFF, RFT, MC

Groups 1 and 2 compare the different effects of the number of followers and the two ratios. Group 3 is used to examine whether firm-followers can also be used to predict comovement. The aim of Groups 4 and 5 is to test the coordination role of both firm-follower and public-follower groups.

Notably, the metrics that we proposed are all simple metrics compared with the metrics selected by previous studies, such as public sentiment. Prior work has shown that simple quantified social media metrics also have a strong impact on firm financial metrics. For example, Luo and Zhang (2013) have revealed the positive relationship between the total page views received by a firm's website and its financial outcome. Another reason that we keep the metrics simple is because they are easy to understand and verify and might attract more practical usage than do complex metrics.

3.3. Prediction model

The left side of Fig. 1 shows the details of the prediction model. It demonstrates that by identifying listed firms' microblogging accounts and retrieving their firm-specific metrics, the model can classify the firms into different homogeneous groups.

[1] We use number of public-followers (P-FO) and number of firm-followers (F-FO) to distinguish the two numbers of followers.

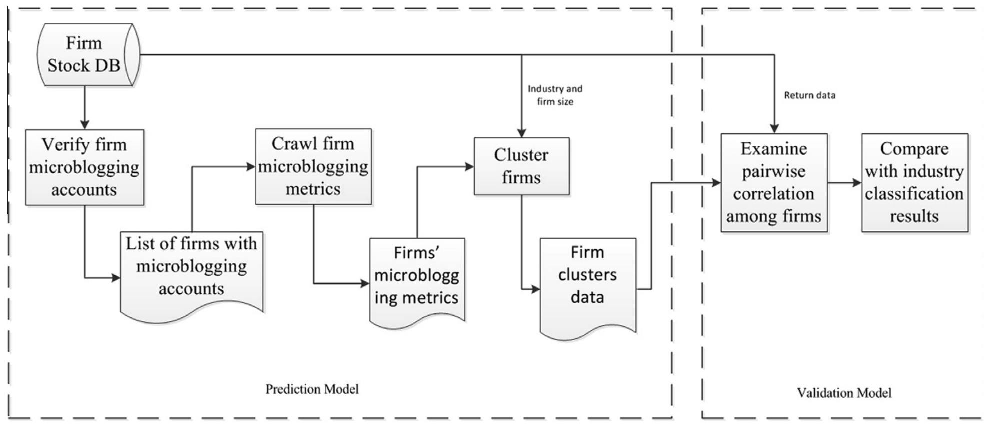

Fig. 1. Prediction and validation model.

The right side of Fig. 1 presents the validation model, which is adopted in the following sections to show the validity the prediction model.

4. Empirical study

4.1. Data collection

To examine the impact of proposed social media metrics on identifying homogeneous stock groups, we used samples of firms that are listed on the NYSE and NASDAQ stock exchanges. We first retrieved the list of the companies that are publicly traded on the two national markets from BvD's database, "OSIRIS: publicly listed companies worldwide". Next, we identified their official Twitter accounts.

We identified 293 American firms that with official microblogging accounts. That number was smaller than expected. Two researchers searched the microblogging platforms using the firms' names separately, and the results were compared and checked by a third researcher. After analyzing the process, we believe the reason that fewer firms than expected have official microblogging accounts is that we adopted rather strict rules about identifying firms' official accounts: (1) The account must be verified. Twitter provides verification of accounts to establish the authenticity of identities. Thus, if a firm has not been verified by Twitter, we treat it as un-official. Many accounts were excluded from our list for this reason. (2) Product or department accounts are also excluded. For example, Microsoft has a number of verified accounts, such as "Microsoft", "Microsoft News", "Microsoft Research", "Microsoft Store", and "Microsoft Cloud", but only the account verified as "Microsoft" is considered its official account. However, some companies do not have a company account, but instead have numerous product or department accounts. For instance, Apple Inc. has "App Store", "Macworld", and a few other product or department focused accounts, but it does not have an account named "Apple". We excluded such cases from our list. This approach ensures that the people who followed companies' accounts are relatively more interested in the companies as a whole than are focused on their products or on a single department.

Although these strict rules exclude some authenticated accounts, it ensured the genuineness of most firms' accounts. Furthermore, although the number of verified accounts is not as high as expected, the number of remains sufficient for our research. In fact, most previous studies' samples were smaller than ours. Many of them only use samples from a single industry (Yu et al., 2013). For instance, Luo et al. (2013) analyze predictive relationships between social media and firm equity value based on 9 high-profile technology firms. Nanda, Mahanty, and Tiwari (2010) use 106 stocks as samples and cluster them into separate groups.

A web crawler program was written to collect data from Twitter using the platform's API. It collected a set of data including the values of the number of followers, number of followings, and number of tweets of each firm account at the parsing time. To generate the number of firm-followers, information about the users followed by each firm was also collected. The program began running on June 30, 2013. Because of the large amount of data, the program ran for five days (crawling ceased on July 4, 2013).

The financial data were collected from Wharton's Center for Research in Security Prices (CRSP) database. We use six-month daily return data from July 1, 2013 to December 31, 2013, to test the predictability of the metrics.

Table 2 shows the descriptive statistics of the data. There is a huge variation in P-FO (Std. Deviation is 917,096.03). The least attractive public traded firm, BLOOMIN' BRANDS, INC., only has 288 followers, whereas, the most attractive firm, Facebook, has nearly 10 million followers. Conversely, the variation in F-FO is much smaller, with a standard deviation of 13.23. This is

Table 2
Descriptive statistics.

	N	Min.	Max.	Mean	Std. deviation
Twitter					
P-FO	293	288	9,596,476	246,800.83	917,096.03
F-FO	293	0	119	10.60	13.23
TW	293	14	161,843	11,033.64	18,643.99
FL	293	2	548,779	7,787.77	34,885.68
RFF	293	.85	115,620.19	833.38	6,937.47
RFT	293	.25	9,353.29	61.34	554.04
MC	293	63,022.38	401,730,071.60	25,676,340.56	52,570,041.55
GICS 25					
P-FO	120	288	8,694,662	262,703.58	892,270.51
F-FO	120	0	119	9.69	14.32
TW	120	14	161,843	12,680.95	20,602.19
FL	120	18	136,495	7,186.55	16,896.12
RFF	120	.85	11,321.17	556.11	1,873.33
RFT	120	.74	384.71	29.79	63.56
MC	120	79,840.80	126,904,330.00	10,773,834.75	20,210,928.28
GICS 30					
P-FO	21	2,969.00	3,448,506.00	225,888.24	742,166.15
F-FO	21	1.00	50.00	8.90	11.19
TW	21	438.00	128,947.00	16,078.43	30,035.89
FL	21	216.00	548,779.00	30,711.19	118,965.36
RFF	21	2.49	479.01	48.95	106.46
RFT	21	1.04	50.39	10.63	11.50
MC	21	283,766.56	244,079,744.20	50,455,035.97	73,842,923.48

TW: the number of tweets; FL: the number of accounts a firm is following.

understandable because there are far fewer firms than general users. Furthermore, as indicated earlier, there are fake accounts on Twitter, but all of the studied firms' accounts are genuine. The variations in the two ratios are also smaller than the basic metrics.

In addition to examining the results based on all firms, we also analyze social media metrics' predictive ability for certain industries. Consumer Discretionary (GICS 25) and Consumer Staples (GICS 30) industries were selected as samples. These two industries include 141 firms, which is nearly half of the 293 firms studied. Furthermore, the two industries are highly consumer related, which may result in them attracting more public attention than other industries.

The variations of basic metrics are also very large in the two industry categories (892,270.51 and 742,166.15, respectively). Interestingly, on average, GICS 30 firms have a much smaller RFF (48.95) than GICS 25 firms (556.11) and all firms (833.38). This is because GICS 30 firms are more willing to follow other accounts. Their mean number of followings is 30,711.19, which is approximately four times as many as GICS 25 (7186.55) and all firms (7787.77). This result indicates that compared to other firms, GICS 30 firms gain fewer followers from each following.

4.2. Comovement identification method

To check whether the proposed metrics groups can be used to predict stock comovement, the stocks were first classified into different homogeneous groups based on the proposed metrics groups. Next, we examined within-group pairwise stock return correlations as an identification of stock comovement.

A K-means clustering algorithm was adopted to classify firms into different groups. Many computer science and information system researchers have adopted such methods to classify stocks. For example, Aghabozorgi and Teh (2014) have proposed a three-phase clustering approach to classifying stocks into different groups based on their historical stock prices. Nanda et al. (2010) have compared three clustering methods and have found that the K-means cluster builds the most compact clusters for stock classification data. The K-means clustering algorithm was selected to classify firms with respect to the metrics groups because it is simple to use but still efficient. Furthermore, the clusters are non-hierarchical and do not overlap.

After classifying stocks into different groups, we need to examine whether the stocks that belong to the same groups have return comovement. In the finance area, a common measurement of stock return comovement involves examining the pairwise stock return correlations (Chan et al., 2007). This paper follows that approach. Supposing that all of the firms are classified into S groups or industries, the within-group/within-industry stock return correlation is calculated as follows.

Let there be M_k firms, $k = 1, \ldots, S$ within each group. We first averaged the pairwise correlations between stock i's return and the return on each of the other members of its group:

$$R_{M_k,i} = \frac{\sum_{j=1, j\neq i}^{M_k} R_{ij}}{M_k - 1}$$

R_{ij} is the time-series correlation between the return on stock i and j.

Next, the average within-group correlation R_{M_k} is calculated as the average of all correlations of the firms within that group:

$$R_k = \frac{\sum_{m=1}^{M_k} R_{M_k,m}}{M_k}$$

Finally, the grand average correlation is given as the average of all groups.

$$R_S = \frac{\sum_{k=1}^{S} R_k}{S}$$

Thus, the within-group correlation and the within-industry correlation can be used to examine whether microblogging metrics can predict stock return comovement.

5. Results

Table 3 shows the within-industry daily return correlations of all-industry stocks ("All"), Consumer Discretionary stocks ("GICS 25") and Consumer Staples stocks ("GICS 30"). For each section, we compared the correlation among three stock groups: the first group ("US") holds all publicly listed stocks, the second group ("Non-Twitter") refers to the stocks without microblogging accounts, and the last group ("Twitter") includes stocks with microblogging accounts.

The results of all-industry stocks and GICS 25 stocks show that except for all-Twitter-stocks, when moving from two-digit to six-digit GICS codes, the correlations increase. This is consistent with the findings of Chan et al. (2007), which also show that firms in a narrower industry category have a higher comovement than those in a broader industry category. However, for US and non-Twitter GICS 30 stocks, the four-digit GICS categories have the lowest correlations. The mechanism that drives this phenomenon remains unknown. GICS 25 stocks are less homogeneous than both all stocks and GICS 30 stocks. The highest correlation of GICS 25 stocks occurs in the Twitter six-digit category (0.250). Most of the GICS 30 categories have lower comovement than the all stocks' categories except the Twitter six-digit category, which equals the highest correlation, Twitter four-digit category (0.324).

In general, Twitter stocks have explicitly higher comovements than US and non-Twitter stocks. The average differences of the correlations among Twitter stocks, US stocks and non-Twitter stocks are 0.106 and 0.116. Although we did not examine whether such effects were merely caused by the act of joining Twitter, it is consistent with the findings of Barberis et al. (2005), which show that when added to a major index, firms experience higher comovement because they may have similar information diffusion efficiency. Twitter, as a popular media platform, has a strong influence on firms' information diffusion, which means that firms that use Twitter versus firms that do not use Twitter may have different media and public exposures. Thus, we believe that the results may have similar causes.

Table 4 depicts the pairwise correlation results of the stock groups clustered by the selected metric groups. K-means clustering may produce different results each time that it runs. Thus, to produce more reliable data, we ran each model 10 times and used the mean of the tests as the result. Furthermore, to control for industry effects, we classify firms based on two- and four-digit GICS industry categories. Each two-digit industry category was set to have six

Table 3
Average Pairwise Correlations between Individual Stocks' Returns of Within-Industry Stocks.

	GICS two-digit	GICS four-digit	GICS six-digit
All			
US	0.168	0.177	0.208
Non-Twitter	0.164	0.171	0.200
Twitter	0.314	0.324	0.323
GICS 25			
US	0.127	0.151	0.175
Non-Twitter	0.116	0.141	0.165
Twitter	0.175	0.211	0.250
GICS 30			
US	0.146	0.142	0.205
Non-Twitter	0.134	0.129	0.186
Twitter	0.263	0.268	0.324

Q1 Please cite this article in press as: Liu, L., et al. A social-media-based approach to predicting stock comovement. *Expert Systems with Applications* (2015), http://dx.doi.org/10.1016/j.eswa.2014.12.049

Table 4
Average Pairwise Correlations between Individual Stocks' Returns of Within-Group Stocks.

Clustered Based on	Group 1	Group 2	Group 3	Group 4	Group 5
Twitter					
GICS two-digit	**0.329**	0.318	0.314	0.320	0.324
GICS four-digit	0.313	**0.338**	**0.326**	**0.338**	0.322
GICS 25					
GICS two-digit	**0.250**	**0.248**	**0.240**	**0.230**	**0.244**
GICS four-digit	0.225	**0.276**	**0.279**	**0.274**	**0.311**
GICS 30					
GICS two-digit	**0.388**	**0.367**	**0.345**	**0.335**	**0.380**
GICS four-digit	**0.370**	0.297	0.313	0.292	0.283

Group 1: FO, MC; Group 2: RFF, RFT, MC;
Group 3: F-Fo, MC; Group 4: F-Fo, FO, MC; Group 5: F-Fo, RFF, RFT, MC.

clusters, and each four-digit industry category was set to have three clusters, to coordinate with the number of GICS six-digit groups. In addition, groups with fewer than three stocks are excluded because the small number of stocks in these groups may cause extreme correlation bias.

The values that are higher than the narrower GICS categories, i.e., if the groups were classified based on GICS two-digit(/four-digit), the compared category is GICS four-digit(/six-digit), are bolded.

Most clustered groups have higher comovement than their corresponding narrower industry categories. GICS 25 stocks, which have the lowest comovement, showed the most promising results. All of their clustered groups are more homogeneous than the narrower industry industries, except for the one classified by P-FO based on the two-digit GICS code. GICS 30 stocks divided based on two-digit GICS codes demonstrated better results than that divided based on four-digit GICS codes. The lowest correlation (0.335) is not only much higher than the four-digit GICS industry correlation of 0.268, but is also greater than the six-digit GICS industry category correlation of 0.324. On the contrary, for Twitter and GICS 25 firms, the metric groups work better based on the four-digit GICS codes. Clustered based on four-digit GICS categories, four out of five metric groups can be used to predict Twitter stocks' comovement, except for Group 5, F-FO, RFF, RFT, and MC, which showed slightly lower comovement than industry categories. Only Group 1 showed a better than industry category result for Twitter stocks when it is used based on two-digit codes. For GICS 25 stocks, groups divided based on four-digit codes are explicitly more homogeneous than the groups divided based on two-digit codes. The smaller number of firms in GICS 30 might be the driver of this phenomenon. There are only 21 GICS 30 firms with Twitter accounts. Thus, when GICS 30 firms were classified based on the four-digit category, some of the clustered groups may have had a very small number of firms, which will cause extreme correlation results.

When designing the models, we selected the ratios and the F-FO as an alternative to P-FO. The results showed no differences between the comovements in the groups clustered by those metrics. This is not as we expected. Groups 2, 3, 4, and 5 were proposed to eliminate the silent influences of P-FO. However, as the results show, there are no differences among the results of the five groups. That finding indicates that despite the simplicity of P-FO, it can predict stock comovement.

Table 5 compares the results of clustered groups and industry categories. It shows the differences between pairwise correlation of within-group and base GICS categories (GICS 2- and 4-digits, respectively). The "Industry" column shows how the correlation increases when moving to a narrower industry category, i.e., the pairwise correlation difference between GICS 4-digit and GICS 2-digit (GICS 2-digit rows) and GICS 6-digit and GICS 4-digit (GICS 4-digit rows). We also calculated the percentage by which the clustered group increased compared to the industry categories and listed the results in parentheses. When moving to the GICS six-digit category, Twitter firms showed a decreased correlation. Thus, when comparing the Twitter groups clustered based on four-digit GICS, we compared them to the differences between four- and two-digit GICS industry categories, 0.010.

The results demonstrated that most clustered groups hugely increased the pairwise correlations compared to the industry categories. In particular, for GICS 30 firms, when clustered based on the two-digit GICS code, the improvements were more than 1000%. The largest improvement was for Group 1, which increased by 2252.76%. GICS 25 firms also showed an average 88.59% improvement over industry categories. Twitter firms showed the least improvement, which only occurred in three groups with an increase of approximately 50%. Such results suggested that the proposed model can be used to predict stock comovement.

6. Conclusion

Comovement is important not only because it plays a vital role in determining portfolio selection strategies (Barberis & Shleifer, 2003; Farrell, 1975; King, 1966) but also because it has an influence on both the market risk premium and the cost of capital (Chelley-Steeley et al., 2013; Roll, 1988). This paper explored the question of whether firm-specific microblogging metrics can be

Table 5
Comparison between Industry Category and Clustered Groups.

Compared with	Industry	Group 1	Group 2	Group 3	Group 4	Group 5
Twitter						
GICS two-digit	0.010	0.015 (50.00%)	0.004 (–)	0.000 (–)	0.006 (–)	0.010 (0.00%)
GICS four-digit	−0.001	−0.011 (–)	0.014 (40.00%)	0.002 (0.00%)	0.014 (40.00%)	−0.002 (–)
GICS 25						
GICS two-digit	0.036	0.075 (108.81%)	0.073 (103.23%)	0.065 (80.91%)	0.055 (53.01%)	0.069 (92.07%)
GICS four-digit	0.039	0.014 (–)	0.065 (66.67%)	0.068 (74.36%)	0.063 (61.54%)	0.100 (156.41%)
GICS 30						
GICS two-digit	0.005	0.125 (2252.76%)	0.104 (1858.52%)	0.082 (1445.52%)	0.072 (1257.79%)	0.117 (2102.57%)
GICS four-digit	0.056	0.159 (183.93%)	0.086 (53.57%)	0.045 (–)	0.024 (–)	0.015 (–)

Group 1: FO, MC; Group 2: RFF, RFT, MC; Group 3: F-Fo, MC; Group 4: F-Fo, FO, MC; Group 5: F-Fo, RFF, RFT, MC.

used to identify homogeneous stock groups and proposed a prediction model with the help of those metrics. The results showed that with simple metrics, social media data can produce better results compared to industry categories.

6.1. Theoretical implications

This paper makes several theoretical contributions. First, most prior studies have focused on overall microblogging metrics, which may result in a high level of many interfering noise. This research identifies firm-specific metrics generated from firms' official microblogging accounts. Second, a great deal of work has concentrated on the predictability of firms' microblogging metrics on stock prices and returns. Although such studies have improved our knowledge of how social media metrics may reflect and predict firms' financial outcomes, this paper sheds light on how firm-specific microblogging metrics have an impact on stock return comovement. To the best of our knowledge, this paper is the first to investigate the impact of firm-specific microblogging metrics on identifying homogeneous stock groups.

The results showed that firms with microblogging accounts have much higher comovements. Although we did not test whether the results were related with whether firms have microblogging accounts, our findings are similar to the results of Claessens and Yafeh (2012), Vijh (1994) and Barberis et al. (2005), who showed that firms in a major index have a higher comovement than firms that are not in the same index. This demonstrates that microblogging platforms, which bypass the traditional information intermediate, have great influence on firms' financial performance.

This research further proved that firm-specific metrics can be anchors for predicting stock return comovement. The results of the models using the number of public-followers, firm-followers, and the two ratios with market capitalization to classify the GICS industry categories into narrower groups demonstrated that within-group correlations are higher than within-industry correlations. In building that theory, this paper also relies on the financial literature of stock return comovement. Its results prove that public behavior can be used to predict stock comovement. This extends information system research and introduces a new approach for financial researchers.

6.2. Practical implications

This study also has implications for financial analysts, investors, firm managers, and policy makers in the financial market. Different from other fundamental comovement factors, social media metrics are current and more flexible. Microblogging metrics are more adaptive to changes, particularly changes related to the firm-followers generated from public firms' social networks. Firm-followers reflect how public firms consider others, which may help investors and financial analysts make decisions. Compared to overall firm social media metrics and sentiment-related metrics, the proposed metrics are more accessible to financial analysts and individual investors. Usually special programs are needed to retrieve overall metrics, whereas most of the firm-specific metrics can be observed and compared directly. Moreover, clustering was based on the GICS industry categories system, which is useful for investors who are only interested in a specific sector. Our results also suggest that investors can use firm-specific microblogging metrics as alternative analysis factors.

The SEC and other regulators should welcome social media as a market risk indicator. Prior research has found that stock return comovement is an identifier of market risk premium (Chelley-Steeley et al., 2013; Roll, 1988). Harmon et al. (2011) have suggested that comovement is the collective behavior characteristic of a public panic and that increasing comovement is significant advance warning of a market crash. Given the ease of monitoring and shorter latency period, regulators can treat social media metrics as a new comovement tester and thus, as a market risk indicator.

6.3. Limitations and future work

Our work has several limitations that can serve as avenues for future studies on social media and firms' financial performance. First, we only focus on microblogging services and do not include other social media channels such as Facebook, blogs, and online discussion boards. Prior research has discussed the overall metrics on these media platforms and their influence on firms' financial outcomes (Gu, Konana, Rajagopalan, & Chen, 2007; Luo et al., 2013; Yu et al., 2013). Future research can aggregate metrics of all of the social media platforms and then examine the impact of those metrics.

Second, future studies may consider the attributes of firms' followers', such as their social activities and their own followers. Some followers may be more active than others, for example, they are more likely to retweet the firms' tweets or to make more comments. Such activities will increase information dissemination, which in turn will influence firms' stock performance.

Third, we adopted very strict rules for identifying firms' official microblogging accounts. Such rules limit the size of our data sample. However, because the sample is sufficient for our research and is larger than many prior studies' samples, more samples might improve our results. There are essentially C_N^2 correlation coefficients when there are N firms in the sample; the number of correlation coefficients increases exponentially as the number of firms increases. Thus, future studies could think of way to identify accounts so that the accuracy of the accounts can be ensured and the size of the samples may be increased.

Moreover, our research has shown that firm-followers have predictability with respect to stock comovement, which indicates that interrelationships between the firms might have an important influence on firms' financial outcomes. Thus, future work should pay attention to utilizing social network analysis techniques in such areas.

Future researchers may be interested in utilizing text and sentiment analysis to explore the impact of the public's behavior and opinions on predicting stock comovement. Our research only focused on the quantitative metrics, but people can retweet and comment on firms' tweets. Adopting sentiment and text analysis in stock comovement analysis may produce a more dynamic result.

Acknowledgments

We thank the editor and anonymous reviewers for their constructive comments, which helped us to improve the manuscript. This research is supported by Major Program of National Natural Science Foundation of China (91218301), Fundamental Research Funds for the Central Universities (JBK120505), and National Natural Science Foundation of China (71401139).

References

Aghabozorgi, S., & Teh, Y. W. (2014). Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applications, 41*(4), 1301–1314.

Alexander, R. M., & Gentry, J. K. (2014). Using social media to report financial results. *Business Horizons, 57*(2), 161–167.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259–1294.

Q1 Please cite this article in press as: Liu, L., et al. A social-media-based approach to predicting stock comovement. *Expert Systems with Applications* (2015), http://dx.doi.org/10.1016/j.eswa.2014.12.049

Aral, S., Dellarocas, C., & Godes, D. (2013). Introduction to the special issue–social media and business transformation: A framework for research. *Information Systems Research, 24*(1), 3–13.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance LXI* (4), 1645–1680.

Barberis, N., & Shleifer, A. (2003). Style investing. *Journal of Financial Economics, 68*, 161–199.

Barberis, N., Shleifer, A., & Wurgler, J. (2005). Comovement. *Journal of Financial Economics, 75*, 283–317.

Bhojraj, S., Lee, C. M. C., & Oler, D. K. (2003). What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research, 41*(5), 745–774.

Bollen, J., Mao, H., & Zeng, X.-j. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Brockman, P., Liebenberg, I., & Schutte, M. (2010). Comovement, information production, and the business cycle. *Journal of Financial Economics, 97*(1), 107–129.

Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics, 70*(2), 223–260.

Chan, L., Lakonishok, J., & Swaminathan, B. (2007). Industry classifications and return comovement. *Financial Analysts Journal, 63*(6), 56–70.

Chelley-Steeley, P., Lambertides, N., & Savva, C. S. (2013). Illiquidity shocks and the comovement between stocks: New evidence using smooth transition. *Journal of Empirical Finance, 23*, 1–15.

Chen, H., De, P., & Hwang, B.-h. (2014). Wisdom of crowds: the value of stock opinions transmitted through social media. *Review of Financial Studies, 27*(5), 1367–1403.

Chou, P.-h., Ho, P.-H., & Ko, K.-C. (2012). Do industries matter in explaining stock returns and asset-pricing anomalies? *Journal of Banking & Finance, 36*(2), 355–370.

Chue, T. K., Gul, F. A., & Mian, G. M. (2014). Investor sentiment and time-varying comovement, working paper.

Claessens, S., & Yafeh, Y. (2012). Comovement of newly added stocks with national market indices: Evidence from around the world. *Review of Finance, 17*(1), 203–227.

Dellarocas, C., Awad, N., & Zhang, X. M. (2004). Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning. In *The 25th international conference on information systems (ICIS)* (p. 34).

Durnev, A., Morck, R., & Yeung, B. (2004). Value-enhancing capital budgeting and firm-specific stock return variation. *The Journal of Finance LIX* (1), 65–105.

Dutt, P., & Mihov, I. (2013). Stock Market Comovements and Industrial Structure. *Journal of Money, Credit and Banking, 45*(5), 891–911.

Elder, J. (2013). Inside a twitter robot factory. *The Wall Street Journal, 12*. URL <http://online.wsj.com/news/articles/SB10001424052702304607104579212122084821400>.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383–417.

Fama, E., & French, K. (1997). Industry costs of equity. *Journal of Financial Economics, 43*, 153–193.

Farrell, J. L. Jr., (1975). Homogeneous stock groupings: Implications for portfolio management. *Financial Analysts Journal, 31*(3), 50–56.

Gallu, J. (2013). SEC approves using facebook, twitter for company disclosures. Bloomberg Report 03 Apr 2013. URL <http://www.bloomberg.com/news/2013-04-02/sec-approves-social-media-use-for-companies-material-disclosure.html>.

Goh, K., Heng, C., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *Information Systems Research, 24*(March 2013), 88–107.

Golbeck, J., Grimes, J., & Rogers, A. (2010). Twitter use by the US congress. *Journal of the American Society for Information Science and Technology, 61*(8), 1612–1621.

Gu, B., Konana, P., Rajagopalan, B., & Chen, H.-W. M. (2007). Competition among virtual communities and user valuation: The case of investing-related communities. *Information Systems Research, 18*(1), 68–85.

Harmon, D., de Aguiar, M. A., Chinellato, D. D., Braha, D., Epstein, I. R., & Bar-Yam, Y. (2011). Predicting economic market crises using measures of collective panic. Available from: <arXiv:1102.2620v>, pp. 1–17.

Jiang, S., Chen, H., Nunamaker, J. F., & Zimbra, D. (2014). Analyzing firm-specific social media and market: A stakeholder-based event analysis framework. *Decision Support Systems, 67*, 30–39.

Kaplan, A. M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons, 54*(2), 105–113.

King, B. F. (1966). Market and industry factors in stock price behavior. *The Journal of Business, 39*(1), 139–190.

Kumar, K., Rajan, R., & Zingales, L. (1999). What determines firm size?, working paper.

Lee, M. K., Shi, N., Cheung, C. M., Lim, K. H., & Sia, C. L. (2011). Consumer's decision to shop online: The moderating role of positive informational social influence. *Information & Management, 48*(6), 185–191.

Lessard, D. (1974). World, national, and industry factors in equity returns. *The Journal of Finance, 29*(2), 379–391.

Liao, S.-H., & Chou, S.-Y. (2013). Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio. *Expert Systems with Applications, 40*(5), 1542–1554.

Liu, L., Li, Q., Xu, Y., & Zhang, Y. (2014). Firm celebrity, reputation and performance: A social media perspective. In *Proceedings of PACIS 2014* (p. 14).

Li, Q., Wang, T., Gong, Q., Chen, Y., Lin, Z., & Song, S.-k. (2014a). Media-aware quantitative trading based on public Web information. *Decision Support Systems*, 93–105.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014b). The effect of news and public mood on stock movements. *Information Sciences, 278*, 826–840.

Luo, X., & Zhang, J. (2013). How do consumer buzz and traffic in social media marketing predict the value of the firm? *Journal of Management Information Systems, 30*(2), 213–238.

Luo, X., Zhang, J., & Duan, W. (2013). Social media and firm equity value. *Information Systems Research, 24*(1), 146–163.

Moskowitz, T., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of Finance, 54*(4), 1249–1290.

Nanda, S., Mahanty, B., & Tiwari, M. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications, 37*(12), 8793–8798.

Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal, 53*(5), 1131–1152.

Phang, C. W., Zhang, C., & Sutanto, J. (2013). The influence of user interaction and participation in social media on the consumption intention of niche products. *Information & Management, 50*(8), 661–672.

Pindyck, R. S., & Rotemberg, J. J. (1993). The comovement of stock prices. *The Quarterly Journal of Economics, 108*(4), 1073–1104.

Rashes, M. S. (2001). Massively confused investors making conspicuously ignorant choices. *The Journal of Finance, 56*(5), 1911–1927.

Rindova, V., Pollock, T., & Hayward, M. (2006). Celebrity firms: The social construction of market popularity. *Academy of Management Review, 31*(1), 50–71.

Rishika, R., Kumar, A., Janakiraman, R., & Bezawada, R. (2013). The effect of customers' social media participation on customer visit frequency and profitability: An empirical investigation. *Information Systems Research, 24*(1), 108–127.

Roll, R. (1988). R^2. *The Journal of Finance, 43*(3), 541–566.

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining – WSDM '12* (pp. 9). New York, New York, USA: ACM Press.

Sprenger, T. O., & Welpe, I. M. (2011). Tweets and peers: Defining industry groups and strategic peers based on investor perceptions of stocks on twitter. *Algorithmic Finance, 1*, 57–76.

Standard & Poor's, (2006). Global industry classification standard (gics) methodology. URL <http://www.unm.edu/~maj/Security%20Analysis/GICS.pdf>.

Sul, H. K., Dennies, A. R., & Yuan, L. I. (2014). Trading on twitter: The financial information content of emotion in social media. In: *Hawaii international conference on system sciences (HICSS)* (p. 10).

Swensen, D. (2009). *Pioneering portfolio management: An unconventional approach to institutional investment*. Free Press.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance LXII* (3), 1139–1168.

Tirunillai, S., & Tellis, G. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science, 31*(2), 198–512.

Vijh, A. M. (1994). S&P 500 trading strategies and stock betas. *The Review of Financial Studies, 7*(1), 215–251.

Wurgler, J. (2000). Financial markets and the allocation of capital. *Journal of Financial Economics, 58*(1–2), 187–214.

Yu, S., & Kak, S. (2012). A survey of prediction using social media. Available from: <arXiv:1203.1647 [cs.SI]>, pp. 1–20.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919–926.

Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets? *Scientific Reports, 4*, 4213.

Zhu, F., & Zhang, X. M. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(March), 133–148.