Social Media Content on Financial Markets

Juheng Zhang

***Abstract*— Stocks are tweeted by investors and are traded in the markets with a potential interplay between daily stock price movements and social media content. We use four daily time-series variables: stock return, volatility, liquidity, and the volume of tweets to study the interdependences and comovements of social media content and stock performance. We find that the Granger causality relationship between the stock liquidity and the volume of tweets over stocks.**

***Index Terms*—stock movements, vector autoregression model, tweets.**

I. Introduction

Social media has increasingly gained popularity in recent years. The study [2] finds that 73% people are active on social media. People use social media sites to share information, read news, and exchange their opinions. The study finds that 84% people trust recommendations of their friends and family more than other sources. Social media play an important role in consumers' decision making. Companies are also adopting social media when devising their marketing strategy or disclosing company news.

The influence of social media in consumer world has also been observed in the financial world. Investors use social media to get market information, company news, traders' opinions, and other investors' sentiment on stock shares [9]. People are influenced by peer recommendations and positive or negative sentiment on social media. The study [8] finds that 62% brokers and traders believe social media content can move stock prices. An infamous case that how social media influences financial markets was the stock market crashed in the year 2015 after the fake tweet posted on Twitter that announced that President Obama was injured in an explosion at the White House. The information was corrected shortly but the impact was devastated with $130 billion in stock value loss and -0.9% S&P500 index drop. This is an extreme example but it shows the huge impact of social media on financial markets.

Stock price changes or any information that may affect stock prices triggers people to tweet on Twitter. Traders and investors post daily price changes of stock picks, and any vital data about the stocks. Stock performance is discussed on social media and people exchange their opinion and share their trading experience on social media. Given 500 million tweets every day, Twitter has a lot of information embedded in it and reaches out to a large number of audience in a few seconds. In this study, we investigate the intertwined relationship between tweets and stock performance. We examine the comovements between tweets and stock performance rather than assuming the tweets as a predictor of stock performance not the other way that stock price movements move the volume of tweets. We believe that tweets influence stock prices, and stock prices trigger people to tweet. We investigate the dynamic relationship of the variables using the time series VAR model.

II. Literature

Existing studies [4, 7] use the tweets to predict the movements of stock market index. They use the OpinionFinder system to find the mood of the tweets and compare it with another algorithm that discovers six different sentiments in the tweets. They find the mood and sentiments are strong predictors of stock market performance. The focus is the collective sentiment of Twitter users and its predictive relationship with the stock market. Other studies [14-19] on social media use different social media metrics to predict firm values or company performance.

We use the vector autoregression (VAR) model to study the dynamic relationship between tweets and stock price movements. The model was proposed by Sims [10] for the interactions of multiple time series variables. The VAR model has been widely applied in different fields [3, 12], and was recently presented to the IS society by Adomavicius et al. [1] and was also adopted in the study [6]. It has great performance in data description/forecasting and structural inferences for multiple times series [e.g., 1, 10, 11, 13]. In the VAR model, each variable is modeled as the lagged values of the variable itself and the lagged values of other endogenous variables in addition to control exogenous variables. This n-variable n-equation model captures the dynamic interdependences and comovements of multiple time series, comparing to the univariate autoregression that is a single variable, single equation model.

III. Data Description

We use the 288 stocks that went public in 2014 as our sample data. We retrieved the accounting numbers of the companies from the Compustat database. Around twenty of the companies have missing data. We provide the basic statistics of the stocks in Table 1. We downloaded the tweets about the stocks during the period from January 1, 2014, to June 1, 2015. We searched the tweets with the stocks' ticker symbols and company names. We added the dollar sign in front of the tick symbol to identify a stock tweet. We also used variant company names when searching the tweets, for example, "Alibaba Group Holding Ltd.," "Alibaba Group Holding," or "Alibaba" for Alibaba company.

Juheng Zhang, Department of Operations and Information Systems, University of Massachusetts Lowell, Lowell, MA, United States,

Table 1: Descriptive Statistics of Stock Prices

Variable	Mean	Std Dev	Min	Max	N
Price	17.24	9.9	4.00	92.7	288
Asset	2305.4	2306.4	2307.4	2308.4	266
Net Income	21.88	363.4	-3426	3750.56	263
Liabilities	1912.1	11108.7	0.131	136959	266
Intangible Asset	216.3	776.5	0	7061	260

We collected detailed information about the company twitter accounts, including the created date, the number of followers, the number of favorites, the number of tweets, etc. Among the 288 sample companies in our study, only 118 had a Twitter account. In Table 2, we provide the statistics for the 118 companies. As shown in the table, the average tweet age was 3.8 years, the maximum was 7.6, and the minimum was -1.3. On average, the companies had 5,667.66 tweets, 36,537 followers, 2,382.07 friends, and 640.75 favorites.

Table 2: Descriptive Statistics of Twitter Accounts

Variable	Mean	Std Dev	Min	Max	N
Age	3.8	2.1	-1.3	7.6	118
Statuses	5667.6	14693.1	0	96155	118
Followers	36537	156683.4	5	1423522	118
Friends	2382.1	8807.8	0	81508	118
Favorites	640.7	2046.8	0	19532	118

IV. Experiment Results

We construct the variable tweets, liquidity, return, and volatility. The variable "tweets" is the number of daily tweets over the stocks, the variable "liquidity" is the number of daily trading volume, "return" is the ratio of close price to open price minus one, and "volatility" is the bid-ask spread as the ratio of the highest bid price to the lowest ask price minus one. We take the natural logs of the variables to eliminate distribution skewness. We aggregate each of the four variables across companies. The aggregation was done with the same weight assigned to each company. Different weights based on the firms' asset sizes are examined in the aggregation and we derived the similar results as when the same weight was used.

We use these information criterion methods and the statistics of these methods such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), or the Hannan-Quinn Information Criterion (HQIC) to select the number of lags in the VAR model. The methods consistently indicate that the appropriate lag length is one. Therefore, the lag length used in the following analysis is one, i.e., $p = 1$.

Data stationarity is an assumption of VAR model. Stationary data have a property with mean, variance, and autocorrelation structure unchanged over time. The Dickey-Fuller unit root test [5] is generally used to check for stationarity. We tested for a unit root with trend and drift, with the results reported in Table 3.

Table 3: Augmented Dickey-Fuller Unit Root Test

	Tweets	Liquidity	Return	Volatility
Intercept	7.9E-12	5.1E-14	0.0761	2.7E-08
Lag 1	8.5E-13	4.2E-14	<2e-16	6.7E-09
Time	0.0142	1.4E-06	0.0859	0.000616
p-value:	2.2E-16	2.2E-16	0.4759	2.2E-16

As shown in Table 3, we can reject the null hypothesis that there is a unit root with a drift or trend in our time series data. The p-values are negligible, with 8.15E-13 on a unit root, 7.91E-12 on drift, and 0.0142 on trend for the variable tweets. Similar results are observed for the variable liquidity and volatility. The results are significant at a 99.9% confidence level.

In the VAR model, the standard practice is to present the results of a Granger causality test, Impulse Response Function (IRF) analysis, and Forecast Error Variance Decomposition (FEVD) analysis. Table 4 provides the p-values for pairwise Granger causality tests. The Granger causality test determines if prior values of a time series variable can help to predict future values of another variable. It is used to examine the causal relationships among endogenous variables. There are several significant Granger causality relationships among the four time series variables. The $tweets_t$ time series significantly Granger-causes the $liqudity_t$ series at a 0.001 level, and the p-value of the Granger-causal relationship is 0.0036. We can conclude that the past values of tweets can predict the future values of trading volume. Other significant Granger causality relationships include $volatility_t \rightarrow liquidity_t$ and $return_t \rightarrow volatility_t$. Recall that the selected lag length is one. The test results suggest that the number of tweets can predict the trading volumes for the next day, and stock returns can predict the next day's bid/ask spread.

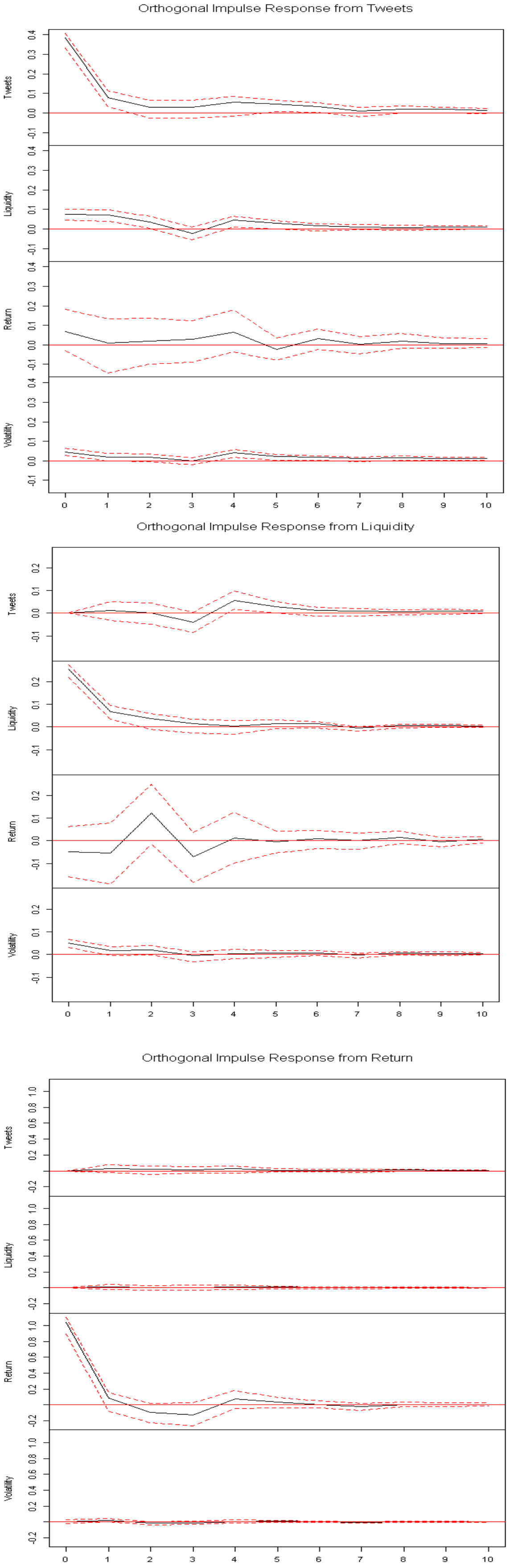
Table 4: Granger Causality Test P Values

	Tweets	Liquidity	Return	Volatility
Tweets	----	0.0036 **	0.747	0.949
Liquidity	0.237	----	0.168	0.232
Return	0.605	0.4681	0.700	0.008**
Volatility	0.624	0.0032 **	0.236	----

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

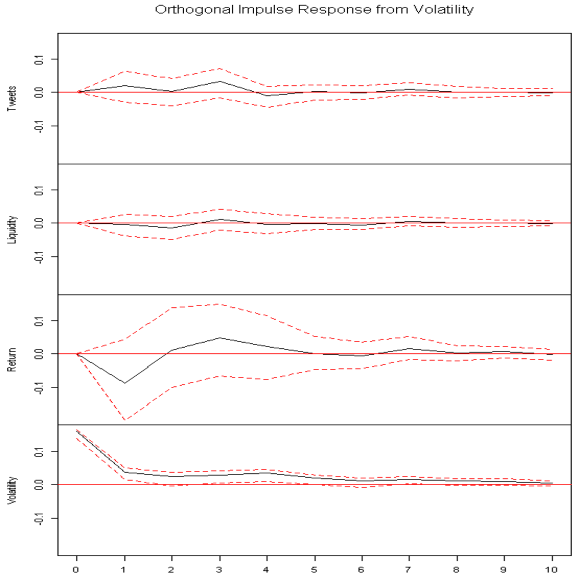
The results of the IRF analysis are graphically represented in Fig.1. The IRF analysis shows the amount of shock on a variable over time due to one unit of change (impulse) in another variable. As demonstrated, the responses from $tweets_t$ on $liquidity_t$ are significantly different from zero over the time periods 0 and 1. The positive responses diminish and converge to zero over time as expected. The results graphically presented in Figure 2 are consistent with the results of Granger causality tests, where trading volume increases as tweets increases, $tweets_t \rightarrow liquidity_t$.

Figure 1: Forecast Error Impulse Responses


Orthogonal Impulse Response from Tweets

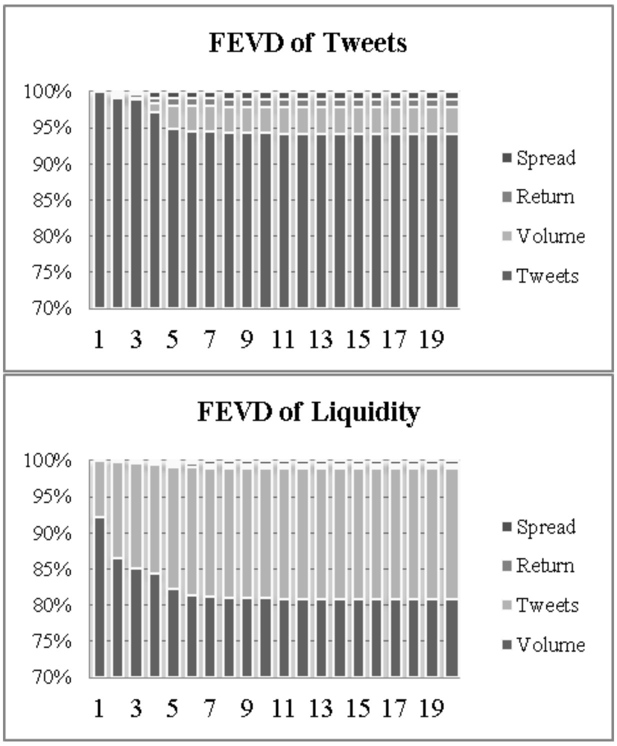
Orthogonal Impulse Response from Liquidity

Orthogonal Impulse Response from Return

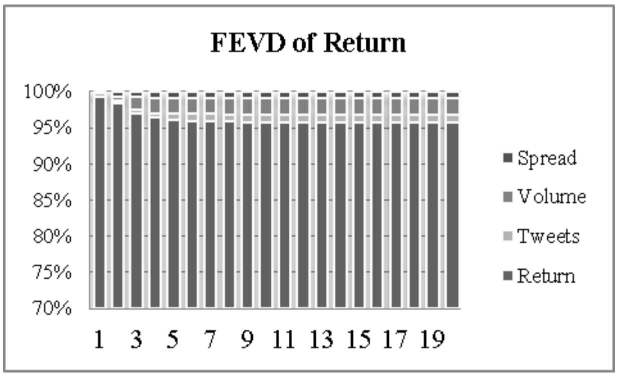

Orthogonal Impulse Response from Volatility

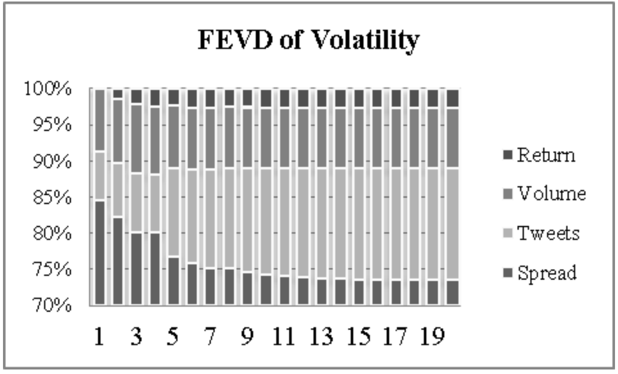
The FEVD analysis determines the amount of forecast error variance of a variable accounted for by shocks to another variable. It indicates the contribution of a variable in explaining another variable. The FEVD analysis results are demonstrated in Fig.2.

Figure 2: Forecast Error Variance Decomposition (FEVD)


FEVD of Tweets

FEVD of Liquidity





The variable $tweets_t$ is responsible for around 20% of the forecast error variance of the $liquidity_t$ variable, and for about 10% of that of the $volatility_t$ variable. Each variable itself accounts for the largest portion of its forecast error variance, as expected. Very little of the error variance of $return_t$ is accounted for by other variables. A possible explanation for the minimal impact of other variables is that while a firm's daily returns are likely to be determined by intrinsic values of the firm such as earnings, $liquidity_t$ and $volatility_t$ are likely to be affected by tweets.

The results of our VAR analysis demonstrate that tweets impacts stock performance and can be used to predict a stock's trading volume (liquidity) in the future. Also, they also show that tweets account for the forecast error variance in the bid/ask spread (volatility) and in trading volumes (liquidity). Stock returns may then, in turn, cause volatility. The four time series: $tweets_t$, $liquidity_t$, $return_t$, and $volatility_t$ are intertwined, and there are significant interdependences between stock performance and tweets about stocks.

V. Conclusion

People chat and comment on stocks in social media. In this study, we ask if tweets have any impact on stocks' price movements and if stock price movements have any influence on tweets. Using daily stock data and tweets over stocks, we employ a VAR model to study interdependences between tweets and stock performance. We find that the volume of tweets Granger-causes stock liquidity can significantly predict the future values of the stock liquidity. The amount of tweets also accounts for the forecast error variance in volatility. There are strong interdependences and comovements between tweets and stock performance.

References

[1]. Adomavicius, G., Bockstedt, J., and Gupta, A. Modeling supply-side dynamics of it components, products, and infrastructure: An empirical analysis using vector autoregression. Information Systems Research, 23, 2 (2012), 397-417.

[2]. Bennett, S. 73% of online adults now use social media. 2013.

[3]. Bernanke, B.S., and Blinder, A.S. The federal funds rate and the channels of monetary transmission. The American Economic Review, 82, 4 (1992), 901-921.

[4]. Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. Journal of Computational Science, 2, 2011 (2011), 1-9.

[5]. Dickey, D.A., and Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74, 366a (1979), 427-431.

[6]. Luo, X., Zhang, J., and Duan, W. Social media and firm equity value. Information Systems Research, 24, 1 (2013), 146-163.

[7]. Mao, H., Counts, S., and Bollen, J. Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051 (2011).

[8]. Milnes, P. Future of funds: How technology and social media are disrupting and opening new opportunities for the fund industry. 2014.

[9]. Shell, A. Wall street traders mine tweets to gain a trading edge. USA TODAY, 2011.

[10]. Sims, C.A. Macroeconomics and reality. Econometrica: Journal of the Econometric Society, 48, 1 (1980), 1-48.

[11]. Stock, J.H., and Watson, M.W. Vector autoregressions. Journal of Economic perspectives (2001), 101-115.

[12]. Walsh, C.E. Monetary theory and policy. MIT press, 2010.

[13]. Watson, M.W. Vector autoregressions and cointegration. Handbook of econometrics, 4 (1994), 2843-2915.

[14]. Zhang, J. Linear discrimination with strategic missing values. Information Systems and Operations Management, Gainesville, FL, USA: University of Florida, 2011.

[15]. Zhang, J. Information revelation and social learning. International Journal of Business and Social Science, 5, 2 (2014), 115-125.

[16]. Zhang, J. Ensuring trust online through the wisdom of crowd. Journal of Internet and e-Business Studies, 2015 (2015).

[17]. Zhang, J. Voluntary information disclosure on social media. Decision Support Systems, 73, 2015 (2015), 28-36.

[18]. Zhang, J., Aytug, H., and Koehler, G.J. Discriminant analysis with strategically manipulated data. Information Systems Research, 25, 3 (2014), 654-662.

[19]. Zhang, J., Khan, R.M., and Shih, D. The rating determinants factored in decision-making for hotel selection. International Journal of Applied Management and Technology, 14, 1 (2015), 1-20.

Juheng Zhang is an assistant professor in the Department of Operations and Information Systems from the Manning School of Business at University of Massachusetts Lowell. She received a Ph.D. in Business Administration from University of Florida. Her research focuses on data analytics and examines information disclosure and manipulation on decision-makings. Juheng Zhang has published in Information Systems Research, Decision Support Systems, and other academic journals.