

High Frequency and Unstructured Data in Finance: An Exploratory Study of Twitter

William Sanger¹, Thierry Warin^{2*}

¹Ph.D. Candidate, Polytechnique Montreal, Canada

^{2*}HEC Montreal, Department of International Business, 3000 Cote-Sainte-Catherine Road Montreal, Quebec, H3T 2A7, Canada

Abstract:

Objective: In this paper, we investigate the question to know whether information spread over Twitter can be useful to design investment strategies on financial markets.

Methods: We compare the influence of two kinds of messages sent on Twitter over two types of returns concerning firms listed on the S&P500. We use logistic-based models to assess the probability of having certain types of returns based on messages published on Twitter.

Results: Financial tweets are positively correlated with higher intraday and overnight returns (1 to 5% returns) while being negatively correlated with lower returns (0 to 1% returns). Non-financial tweets are not significantly related to such returns.

Conclusion: From a practical standpoint, investment strategies could be designed following these findings to optimize some gain opportunities depending on the investment day, the targeted industry and live activity on Twitter.

Keywords- Social media, Stock prices, Big data, S&P500, Twitter, High frequency, Unstructured data

INTRODUCTION

“Breaking: Two Explosions in the White House and Barack Obama is injured.” (The Associated Press, 10:07, April 23rd 2013) 72 characters later, the S&P500 index lost more than 121 billion US dollars until the tweet was proven false and confirmed to be sent by a hacked Associated Press account. A 1.68 billion-dollar bill for each character written. However, this particular event shed light on the implication of spreading news on the stock markets, especially through social media.

Our research question is to know whether information running through Twitter explain some of the stock price variations on the S&P500. Twitter is a new way of spreading information, not only through short sentences, but also by allowing users to emphasize one particular piece of information by retweeting. As such, is Twitter a complement to traditional ways of spreading news, or is it a revolution? Applied to a particular market - the financial market - this question takes a whole different dimension. Indeed, information is at the core of finance. In theory, there is no way to beat the market return. However, when one investor has more information than another one, then she can beat the market. In this regard, can we extract some extra information from messages on Twitter that can lead to some investors beating the market all the time?

Since the advent of modern finance, information has taken a central role in every mechanism of the industry. While Markowitz established the theoretical framework upon which relies the Capital Asset Pricing Model, strong assumptions have been made. Among them figures the fact that an investor will act in a rational way in order to maximize her returns while minimizing her risks [1]. Twenty years later, [2] characterized the market efficiency regarding information: there should be no gain opportunity on the stock markets because they are defined by random walk patterns, while all information is known at any time because prices instantly reflect new events.

These two assumptions have shown their limitations, especially

during financial crises (October 97, the Internet bubble or the US financial crisis of 2007 for naming a few examples). If crises cannot be explained by rational thinking à la Markowitz, perhaps the emotional involvement of people is a potential answer. Behavioral finance interprets financial markets as a proxy to reflect the social mood [3]: crises could be provoked by what [4] describe as “animal spirits”.

With the advent of social media and the democratization of low-cost and efficient informatics systems, the wisdom of crown has never been so accessible. Every day, 500 millions tweets are publicly sent around the world. More than a billion users are connected on Facebook. And this is only the tip of the iceberg. Big Data is considered as one of the most promising futures for finance, especially for risk management [5]. By using a robust framework inherited from the financial industry, [6] developed dashboards to manage risks based on social media scrutiny. We define Big Data as a mix of structured and unstructured data (stock prices, heartbeats and pictures for example), produced in real-time as part of longitudinal data [7,8]. Highlighted the fact that qualitative data can reflect information that is not optimally incorporated into stock prices. In that sense, Twitter messages fit the definition of unstructured data massively produced and the social media became the subject of several studies in finance and social sciences.

While many have used sentiment analysis technics or machine learning approaches, few econometric analyses have yet been published. “Twitter hedge funds” were highly mediatized but did not earn the expected returns of the studies. Here lies our challenge: how to assess efficiently the massive flow of information for the financial markets? The main objective of this research is to quantify investment opportunities following an increase in Twitter messages depending on (1) weekdays and (2) industry types. We expect that intraday and overnight returns could be correlated with the volume of tweets being published. To summarize, our overall research question is to know whether new forms of information diffusion such as Twitter can add to the current financial information flows and explain some

changes in stock returns. This leads to two sub-questions and two hypotheses:

RQ1. Is there a difference between tweets exchanged after the markets are closed and tweets written during the day, when markets are open, in explaining stock price changes? For that matter, we use two dependent variables: intraday and overnight returns.

RQ2. Is there a difference between tweets written by financially literate people (using company tickers) and tweets written by the layman (using company names) in explaining stock price changes?

H1. The first hypothesis is related to the first research question in the sense that we assume that professionals tweet after they are done at work, and their comments may be interesting for investment decisions implemented on the next day.

H2. The second hypothesis is related to the second research question in the sense that we assume that tweets written by professionals (using tickers) provide more useful financial information than tweets written by the layman (using company names).

The rest of the paper is organized as follows: section (2) presents the literature review concerning the use of Internet as a source of information for the financial markets, including forums, search engines and social media in the later years. In section (3), we describe our dataset composed of stock prices and the number of messages sent on Twitter about 71 firms of the S&P500. We use logistic-based estimations, which are detailed in section (4), while results are interpreted in the last section of the paper.

LITERATURE REVIEW

Forums and Blogging Websites

In early 2000, financial blogs were used to discuss about stock performances [9] studied Yahoo! Finance comments to understand the characteristics of the most spoken firms on the website. He found that when the number of messages doubles overnight, the following daily return is on average 0.18% higher. RagingBull.com, one of the first financial forums, was also the subject of a few studies. Through event studies, the impact of messages cannot be anticipated by more than a day in advance [10]. Trading volumes have also been correlated to the number of messages written on the corresponding companies and thus Internet messages cannot be considered as noise [11]. By investigating topics shared on Engadget.com, a technological blog, [12] were able to predict stock performances' magnitude in 78% of the cases (and 87% of the returns' sign). Propagating rumors on the Internet have been linked to trading volumes using HotCopper.com website [13].

A second part of the studies regarding forums and blogging websites focuses on the concept of opinion leadership. It is opposed to the mass' conformity and how false information interferes with the wisdom of crowd: more precisely, [14] interpret noise as distortion of a signal. An opinion leader is able to influence the perception of other ones by accentuating his own positions to balance the public's inertia [15]. This capacity is a key element towards understanding the importance of opinion leaders inside a network. Identifying how connected a person is and how novel the information transmitted is determine if a person can be considered as an opinion leader on the Internet [16]. These results can be used for predicting potential sales of

product or how a person's network will be inclined to buy a similar object [17,18]. Finally, [19] differentiate two behavior of influencers: "agitator", which is a person stimulating discussions and "summarizer", which is a person trying to give a clear picture of a situation.

Search Engines

Search engines, by being the medium through which users access information on the Internet, have become a valuable source of financial data. Google has released Google Trends, a tool aggregating the volume of search queries to interpret its users' behavior. Several billion queries are made every day, and thus represent a huge opportunity in terms of Big Data analysis [20]. For example, relationships linking search queries and car sales have been [21].

In Finance, trading volumes of S&P500's firms have been positively correlated to search queries through Google Trends [22]. This result was confirmed with other stock markets (Dow Jones, CAC40, DAX and FTSE). By adding Google Trends data, predictive models have shown more accurate explanation capacities. However, search queries and stock volatility seems to be mutually influenced, since "the investors' attention to the stock market rises in times of strong market movements [and] a rise in investors' attention is followed by higher volatility" [23]. In fact, search engines also appear to reduce the information gap between investors: search engines could be used as a proxy for "naïve" investors' behavior [24]. Only a few companies are widely looked on the Internet, but this non-expert point of view could be interpreted as the wisdom of crowd regarding stock markets [25].

Besides trading volumes, stock returns have tried to be anticipated using a similar methodology. Sudden announcements of firms with high growth and a narrow products' offer are anticipated using Google Trends [26,27] found that more queries about a company leads to an increase in trading volumes, but also a decrease of the expected returns. However, when maximum (minimum) yearly values of stock prices have been reached, the predictive power of Google Trends is increased (decreased). Chinese stocks' abnormal returns were studied using a similar tool, Baidu Index, which counts the raw number of queries made on Baidu' servers [28].

The main limitation of Google Trends lies in the results given to its users: only aggregated data is provided, on a weekly level. Social media, by providing real-time and open information, could become a more appropriate tool for financial predictions [29].

Social Media

Even though Facebook gathers more than a billion users, only a few studies were made in finance [30] used the social network as a proxy for accessing the emotion of users. He created an index called Facebook's Gross National Happiness (GNH), which takes into account updates from users' status. A standard deviation of the GNH is correlated with an increase of stock returns the next day. In a now famous study, [31] showed that emotions spread between users and could be manipulated on the social network. Ethical questions aside, these results reinforce the notion of a highly connected network that could be influenced by its users, alike on the financial markets with herd behaviors.

eToro was part of several studies released by MIT's Media Lab. Described as a "fun and accessible" way to democratized trading [32], the social network was able to shed light on how likely an information could become trending [33]. Also, researchers found that the reputation of a user is not due to his past trading performances but more to his links with other users [34].

Since its release in 2006, Twitter is the subject of studies in various fields, including election outcomes, crisis management or disease outbreaks. Few rules regulate the messages sent on the social network: message cannot be longer than 140 characters and can be referenced by the use of a "hashtag" ("#") in order to facilitate search regarding to a particular topic. For financial purposes, using the ticker symbol preceded by the "\$" symbol is commonly adopted (i.e. \$AAPL in order to write a message about Apple, or \$NFLX regarding Netflix).

In order to process the amount of messages sent every day, several technics have been employed, and in particular sentiment analysis [35] investigates the predictive power of emotion through tweets by using a 6-level categorization method. They found that messages associated to the emotion "calm" could influence stock returns, and thus results could be anticipated as far as 4 days in advance. Stock performances (returns and trading volumes) are also linked to Twitter metrics such as sentiment or message volume [36]. Stock prices seem not to fully reflect the information publicly available, even though the information gap tend to be quickly filled [37]. In a detailed econometric study, [38] could not reproduce [35] impressive results but have been able to predict 70% of the time the DJIA's direction, 58.08% of the time the NASDAQ's direction and 68.63% of the time the S&P500's outcomes.

Due to the massive adoption of social media by Internet users, firms have never been so exposed to the "buzz". Adopting a game-theory approach, [39] illustrated the importance of taking into account words spread on the Internet and how to react to a reputation crisis. A message could become out of control and thus harm the firm's stock price [40]. However, researchers have also use Twitter as an opportunity for trading activities since the information is not instantly incorporated into prices [41,42]. Derwent Absolut Return, commonly called "Twitter hedge fund" at its beginning, used the methodology developed by Bollen and al. but could not transform their theoretical results into practical success [43].

This paper relies on a methodology inherited by the studies using search engines, which is a volume-based methodology instead of a sentiment analysis methodology. Our hypothesis stipulates that investment strategies could be implemented in order to maximize an investor's gains on the stock market.

DATA

Structured Data: Stock Information

Financial data are from Yahoo! Finance database. We extracted the following information regarding the 500 firms composing the S&P500: opening price and closing price. These metrics were established on a daily basis, for 251 trading days, starting on May 1st 2012 and ending on May 1st 2013.

From these data, we define 2 types of return:

$$\text{Intraday.return}_i = \frac{\text{Closingprice}_i - \text{Openingprice}_i}{\text{Openingprice}_i} \quad (1)$$

$$\text{Overnight.return}_i = \frac{\text{Openingprice}_i - \text{Closingprice}_{i-1}}{\text{Closingprice}_{i-1}} \quad (2)$$

We create dummy variables in order to control for weekdays (Monday to Friday) and industry types following the Global Industry Classification Standard (Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Telecommunication Services, Utilities).

Unstructured Data: Tweets

Two types of messages were collected regarding the same firms. (1) We collected the number of time a company was named on Twitter per day (i.e. "Google" for Google Inc. or "Microsoft" for Microsoft Inc.). (2) The second type of tweet is called financial tweets, characterized by the presence of the "\$" symbol before the ticker of a listed firm (i.e. "\$GOOG" for Google Inc. or "\$MSFT" for Microsoft Inc.). Data regarding these two types were aggregated by day and also collected from May 1st 2012 and May 1st 2013 using the People Browsr website. For understanding purposes, the first type of tweets is later called Name and then Name.Index, while the second type of tweets is later called Ticker and then Ticker.Index.

From the 500 companies of the S&P500, we selected 71 of them in order to only keep firms that have on average 30 financial tweets per day. The list of queries used for the selected firms is provided on Tables 1a & 1b.

In Table 2, we provide some descriptive statistics for each variable.

Finally, we test the correlation coefficients between the variables and provide the correlation matrix in Table 3.

METHODOLOGY

Our research question is to know whether tweets can explain some of the variation in stock prices. In other words, is Twitter one of the channels of financial information? Is it an evolution or a revolution in finance? It is particularly interesting to study Twitter and its impact on the financial market, because Twitter (1) provides short messages forcing users to be to the point, (2) is real-time, (3) allows the users to emphasize a message by retweeting (hence creating a buzz) and (4) allows the users to tweet only about financial products by using a \$ sign before the company name for instance. The latter point is of particular interest from a statistical perspective, because it allows us to have a great sample, covering almost all of the population tweeting about finance.

Again, our research questions are as follows:

RQ1. Is there a difference between tweets exchanged after the markets are closed and tweets written during the day, when markets are open, in explaining stock price changes? For that matter, we use two dependent variables: intraday and overnight returns.

RQ2. Is there a difference between tweets written by financially literate people (using company tickers) and tweets written by the layman (using company names) in explaining stock price changes?

H1. The first hypothesis is related to the first research question in the sense that we assume that professionals tweet after they are done at work, and their comments may be interesting for investment decisions implemented on the next day.

H2. The second hypothesis is related to the second research question

ID	Company	Name	Ticker	Industry
1	Agilent Technologies	agilent	A	Healthcare
2	Alcoa	alcoa	AA	Materials
3	Amazon.com	amazon	AMZN	Consumer Discretionary
4	American International Group	american intl group	AIG	Financials
5	Apple	apple	AAPL	Information Technology
6	AT&T	at&t	T	Telecommunications Services
7	Bank of America	bank of america	BAC	Financials
8	Best Buy Co.	bestbuy	BBY	Consumer Discretionary
9	Boeing	boeing	BA	Industrials
10	Caterpillar	caterpillar	CAT	Industrials
11	Celgene	celgene	CELG	Healthcare
12	Chesapeake Energy	chesapeake	CHK	Energy
13	Chevron	chevron	CVX	Energy
14	Chipotle Mexican Grill	chipotle mexican	CMG	Consumer Discretionary
15	Cisco Systems	cisco	CSCO	Information Technology
16	Citigroup	citigroup	C	Financials
17	Cliffs Natural Resources	cliffs natural	CLF	Materials
18	Coach Inc	coach inc	COH	Consumer Discretionary
19	Coca-Cola	coca cola	KO	Consumer Staples
20	Dell	dell	DELL	Information Technology
21	Dominion Resources	dominion resources	D	Utilities
22	Dow Chemical	dow chemical	DOW	Materials
23	eBay	ebay	EBAY	Information Technology
24	Exxon Mobil	exxon	XOM	Energy
25	First Solar	first solar	FLSR	Industrials
26	Ford Motor	ford	F	Consumer Discretionary
27	Freeport-McMoran	freeport memoran	FCX	Materials
28	General Electric	general electric	GE	Industrials
29	General Motors	general motors	GM	Consumer Discretionary
30	Goldman Sachs	goldman sachs	GS	Financials
31	Google	google	GOOG	Information Technology
32	Hewlett-Packard	hewlett	HPQ	Information Technology
33	The Home Depot	home depot	HD	Consumer Discretionary
34	Intel Corporation	intel	INTC	Information Technology
35	International Business Machines	ibm	IBM	Information Technology
36	Johnson & Johnson	johnson & johnson	JNJ	Healthcare
37	JPMorgan Chase & Co	jpmorgan	JPM	Financials
38	Kellogg Company	kellogg	K	Consumer Discretionary
39	Eli Lilly & Company	eli lilly	LLY	Healthcare
40	Loews Corporation	loews	L	Financials
41	Lowe's Companies	lowe's	LOW	Consumer Discretionary
42	Macy's	macy's	M	Consumer Discretionary
43	MasterCard	mastercard	MA	Information Technology
44	McDonald's	mcdonald	MCD	Consumer Discretionary
45	Merck & Co	merck	MRK	Healthcare
46	Microsoft Corporation	microsoft	MSFT	Information Technology
47	Monster Beverage Corporation	monster beverage	MNST	Consumer Staples
48	Morgan Stanley	morgan stanley	MS	Financials
49	Netflix	netflix	NFLX	Information Technology
50	Nike	nike	NKE	Consumer Discretionary
51	NVIDIA Corporation	nvidia	NVDA	Information Technology
52	Oracle Corporation	oracle	ORCL	Information Technology
53	J.C. Penney Corporation	penney	JCP	Consumer Discretionary
54	Pfizer	pfizer	PFE	Healthcare
55	The Priceline Group	priceline.com	PCLN	Information Technology
56	Procter & Gamble	procter gamble	PG	Consumer Staples
57	Qualcomm	qualcomm	QCOM	Information Technology
58	Ryder System	ryder system	R	Industrials

59	Salesforce.com	salesforce	CRM	Information Technology
60	Starbucks Corporation	starbucks	SBUX	Consumer Discretionary
61	Target Corp.	target corp	TGT	Consumer Discretionary
62	TECO Energy	teco energy	TE	Utilities
63	TripAdvisor	tripadvisor	TRIP	Consumer Discretionary
64	United States Steel Corp.	united states steel corp	X	Materials
65	Verizon Communications	verizon	VZ	Telecommunications Services
66	Visa Inc.	visa inc	V	Information Technology
67	Wal-Mart Stores	walmart	WMT	Consumer Staples
68	The Walt Disney Company	disney company	DIS	Consumer Discretionary
69	Wells Fargo & Company	wells fargo	WFC	Financials
70	Yahoo! Inc.	yahoo	YHOO	Information Technology
71	Yum! Brands	yum!	YUM	Consumer Discretionary

Table 1: (A): (ID 1-35): list of the 71 selected firms for the study. (B): (ID 36-71): list of the 71 selected firms for the study.

Variable	Ticker	Name	Overnight Return	Intraday Return
Number of obs.	16043	16043	16043	16043
Average	144.5	16732	0.00012	0.00051
Median	54	949	0.003	0.00049
Max.	19546	1091212	0.3944	0.4499
Min.	0	0	-0.754	-0.1594
Std. Dev.	488	44564	0.0143	0.0159
1 st Q.	31	55	-0.0035	-0.0067
3 rd Q.	112	7517	0.004	0.0078

Table 2: Descriptive statistics.

Variables	Intraday Return	Overnight Return	Ticker	Name
Intraday Return	1	-0.01	-0.009	-0.004
Overnight Return	-0.01	1	-0.008	0.007
Ticker	-0.009	-0.008	1	0.526
Name	-0.004	0.007	0.526	1

Table 3: Correlation matrix.

in the sense that we assume that tweets made by professionals (using tickers) provide more useful financial information than tweets written by the layman (using company names).

From an econometric perspective, the methodology used for this study relies on logistic estimations. Logistic models evaluate the impact a variable over the probability that the studied variable changes states. The dependent variable is reduced to a binary variable that can be adapted to different models. In our case, we evaluate the two different returns and implemented three models concerning the magnitude of these returns, such as:

$$Model_1 = \begin{cases} 1 & \text{if return is positive} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$Model_2 = \begin{cases} 1 & \text{if } 0 < return \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$Model_3 = \begin{cases} 1 & \text{if } 1 < return \leq 5 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

In order to assess the effect of Twitter messages on firms, we control for two different types of fixed effects (weekdays on the one hand, and industry types on the other hand). This allows us to evaluate the specific impact of an increase (decrease) of tweets depending on the weekday or the kind of industry, such as:

$$Pr(\rho_{t,\phi}) = \alpha_0 + \alpha \cdot \rho_{t-1} + \beta \cdot \tau_{t-1} + \gamma \cdot \omega_t + \delta \cdot \sigma_t + \varepsilon_t \quad (6)$$

Where $\phi = \{Model_1; Model_2; Model_3\}$ is the model used for the selected return $\rho_{t,\phi} = \{intraday; overnight\}$ $\tau_{t-1} = \{\text{ticker; name}\}$ the number of tweets published about a company mentioning either its ticker or its name the day before the studied return, the ω_t weekdays and σ_t a value corresponding to the Global Industry Classification Standard.

In terms of methodology, we compute the log-odds, odds-ratios and predicted probabilities of our models, as well as testing the validity of the impact of our variables by a Wald-test. Finally, in our logistic estimations, we use Monday and the Energy sector as our reference points.

RESULTS

We provide a summary of the logistic estimations' results in the following tables. Again, results have to be interpreted relatively to the reference points. We should also pay attention to the low R-squared values. They should be considered in the context of an exploratory study. Also, although the values are low, they have to be interpreted as follows: they represent how much we can add to the level of information already available in the financial markets.

Overnight returns and financial tweets

In Table 4, we compare the influence of financial tweets regarding to the three models of overnight returns. Firstly,

Type of tweet: name Type of return: Overnight	Model 1: 0% < Return			Model 2: 0% < Return < 1%			Model 3: 1% < Return < 5%		
	Log-odds		Odds-ratios	Log-odds		Odds-ratios	Log-odds		Odds-ratios
Constant	-0.4772	***	0.6205	-0.484	***	0.6162	-3.321	***	0.0361
Overnight return t_{-1}	-4.634	***	0.0097	-1.639		0.1941	-4.478	***	0.0114
Ticker t_{-1}	-3.332 x 10^{-5}		0.9999	-2.04 x 10^{-4}	***	0.997	1.846 x 10^{-4}	***	1.0002
Weekdays (ref: Monday)									
Tuesday	0.7058	***	2.0254	0.714	***	2.0421	0.12		1.1275
Wednesday	0.3643	***	1.4395	0.3518	***	1.3713	0.2145	**	1.2392
Thursday	0.5476	***	1.7291	0.4823	***	1.6197	0.3149	***	1.3702
Friday	0.4121	***	1.0599	0.276	***	1.3178	0.4379	***	1.5495
Industrial Sectors (ref: Energy)									
Materials	0.1503		1.1621	2.986 x 10^{-2}		1.0303	0.3605	*	1.434
Industrials	0.03079	***	1.3605	-9.266 x 10^{-3}		0.9907	0.9684	***	2.6337
Consumer Discretionary	-1.842 x 10^{-3}		0.9981	5.545 x 10^{-2}		1.0570	0.5003	**	0.6063
Consumer Staples	0.12	*	1.1286	3.545 x 10^{-2}		1.0035	0.4493	***	1.5673
Health Care	0.1699	**	1.1851	-0.1823	**	0.8333	1.084	***	2.9563
Financials	0.1318	**	1.1408	6.29 x 10^{-2}		1.0649	0.4423	***	1.5563
Information Technology	7.521 x 10^{-2}		1.0781	0.1412		1.1516	-0.4533		0.6355
Telecommunication Services	0.1192		1.1265	0.292	***	1.3390	-1.07	***	0.343
Utilities	0.1181		1.1253	-0.3185	***	0.7272	1.337	***	3.8073
Number of observations	16043			16043			16043		

P-value levels: *** < 0.01; ** < 0.05; * < 0.1; - < 0.15

Table 4: Log-odds and odds-ratios of having an overnight return higher than 0% (Model 1), between 0 and 1% (Model 2) or between 1 and 5% (Model 3) depending on financial tweets (Ticker), weekdays and industrial sectors.

Type of tweet: ticker Type of return: overnight	Model 1: 0% < Return	Model 2: 0% < Return < 1%	Model 3: 1% < Return < 5%
	Predicted Probabilities	Predicted Probabilities	Predicted Probabilities
Industrial Sectors			
Energy	0.4705	0.3751	0.0437
Materials	0.5080	0.3821	0.0616
Industrials	0.5473	0.3729	0.1076
Consumer Discretionary	0.4700	0.3881	0.0270
Consumer Staples	0.5007	0.3759	0.0669
Health Care	0.5129	0.3334	0.1192
Financials	0.5034	0.3899	0.0665
Information Technology	0.4893	0.4087	0.0282
Telecommunication Services	0.5003	0.4456	0.0154
Utilities	0.5000	0.3038	0.1484
Overnight_return t_{-1}	0.00019	0.00019	0.00019
Ticker t_{-1}	128.	128.	128.
Weekdays	Wednesday	Wednesday	Wednesday
Number of observations	16043	16043	16043

Table 5: Predicted probabilities of having an overnight return holding all variables (Ticker t_{-1} , Overnight_return t_{-1} , Weekdays) at their meaning.

Ticker t_{-1} is highly significant (as confirmed by a Wald-test on this variable) but its effect on the probability of having such returns are opposite. In fact, for a unit increase in Ticker t_{-1} , the log-odds of having an overnight return between 0 and 1% decrease by 2.04×10^{-4} but for the third model (from 1 to 5%) the log-odds increase by 1.846×10^{-4} .

Secondly, holding all variables at their means (Ticker t_{-1} equals to 127 tweets, Weekdays being Wednesday and Overnight_return t_{-1} being 0.00012), the predicted probabilities of having the three different overnight returns can be computed based on the industry classification (Table 5). For the first model,

Industrials (54%), Health Care (51.3%) and Materials (50.8%) are the industry types with the highest probabilities of having such returns; for the second model, it is Telecommunication Services (52.4%), Information Technology (48.7%) and Consumer Discretionary (46.5%); for the third model, it is Utilities (14.8%), Health Care (11.9%) and Industrials (10.8%). Thirdly, we can compare these predicted probabilities for different level of tweeting activity. Figure 1 provides such analysis, with results depending on (1) the day of the week (1 to 5), (2) the type of industry (1 to 10) and (3) the number of financial tweets published (from 0 to 10,000). Here, the more a

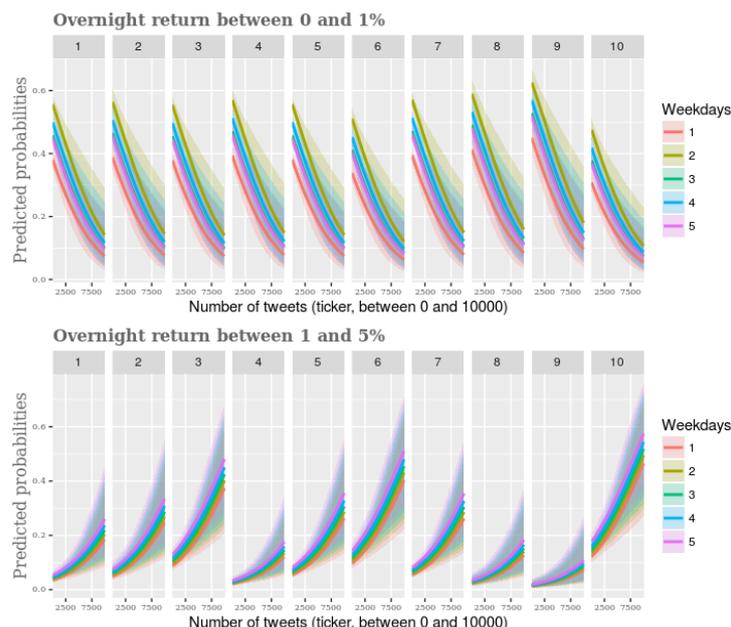


Figure 1: Predicted probabilities of having an overnight return between 0 and 1% (upper panel) and between 1 and 5% (lower panel) depending on financial tweets (between 0 and 10000), on weekdays (1 – Monday, 2 – Tuesday, 3 – Wednesday, 4 – Thursday and 5 – Friday) and on industry sectors (1 – Energy, 2 – Materials, 3 – Industrials, 4 – Consumer Discretionary, 5 – Consumer Staples, 6 – Health Care, 7 – Financials, 8 – Information Technology, 9 – Telecommunication Services and 10 – Utilities).

Type of tweet: name Type of return: overnight	Model 1: 0% < Return			Model 2: 0% < Return < 1%			Model 3: 1% < Return < 5%		
	Log-odds		Odds-ratios	Log-odds		Odds-ratios	Log-odds		Odds-ratios
Constant	-0.4753	***	0.6205	-0.4817	***	0.6177	-3.323	***	0.0360
Overnight_return _{t-1}	-4.655	a*	0.0097	-1.635		0.1949	-4.673	***	0.0093
Name _{t-1}	-5.823 x 10 ⁻⁷	-	0.9999	-2.870 x 10 ⁻⁷		0.9999	1.488 x 10 ⁻⁶		1.0000
Weekdays (ref: Monday)									
Tuesday	0.7016	**	2.0254	0.6983	***	2.0102	0.1321		1.1412
Wednesday	0.3591	***	1.4395	0.2966	***	1.3453	0.2309	**	1.2597
Thursday	0.5417	***	1.7291	0.4617	***	1.5868	0.3383	***	1.4025
Friday	0.4069	***	1.5099	0.2573	***	1.2934	0.4544	***	1.5751
Industrial Sectors (ref: Energy)									
Materials	0.1501		1.1621	3.385 x 10 ⁻²		1.0344	0.3548	*	1.4259
Industrials	0.3051	***	1.3605	-2.338 x 10 ⁻²		0.9768	0.9805	***	2.6658
Consumer Discretionary	-9.791 x 10 ⁻³		0.9981	6.130 x 10 ⁻²		1.0632	-0.5242	**	0.5920
Consumer Staples	0.1096	*	1.1286	6.873 x 10 ⁻³		1.0068	0.4230	***	1.5265
Health Care	0.1691	**	1.1851	-0.1824	**	0.8332	1.083	***	2.9522
Financials	0.1007		1.1408	3.275 x 10 ⁻²		1.0332	0.4316	***	1.5396
Information Technology	7.119 x 10 ⁻²		1.0781	0.1358		1.1455	-0.4550		0.6344
Telecommunication Services	0.1202		1.1265	0.2971	***	1.3459	-1.075	***	0.3413
Utilities	0.1177		1.1253	-0.3214	***	0.7251	1.340	***	3.8178
Number of observations	16043			16043			16043		

p-value *** < 0.01; ** < 0.05; * < 0.1; - < 0.15

Table 6: Log-odds and odds-ratios of having an overnight return higher than 0% (Model 1), between 0 and 1% (Model 2) or between 1 and 5% (Model 3) depending on non-financial tweets (Name), weekdays and industrial sectors.

company is tweeted about, the less it is likely to have an overnight return between 0 and 1% (while it has higher chances of having an overnight return between 1 and 5%). Finally, weekdays and industry types also influence these predicted probabilities.

Overnight returns and non-financial tweets

Our hypothesis stipulates that financial tweets (Ticker) provide

more information than non-financial tweets (Name). Name_{t-1} has a p-value of 0.14 for Model 1 and is not significant for Model 2 (Table 6). The Wald-test confirms these results, while Name_{t-1} is significant for Model 3. Also, the magnitude of the log-odds has to be noticed between financial and non-financial tweets. Again, holding all variables at their means (Name_{t-1} equals to

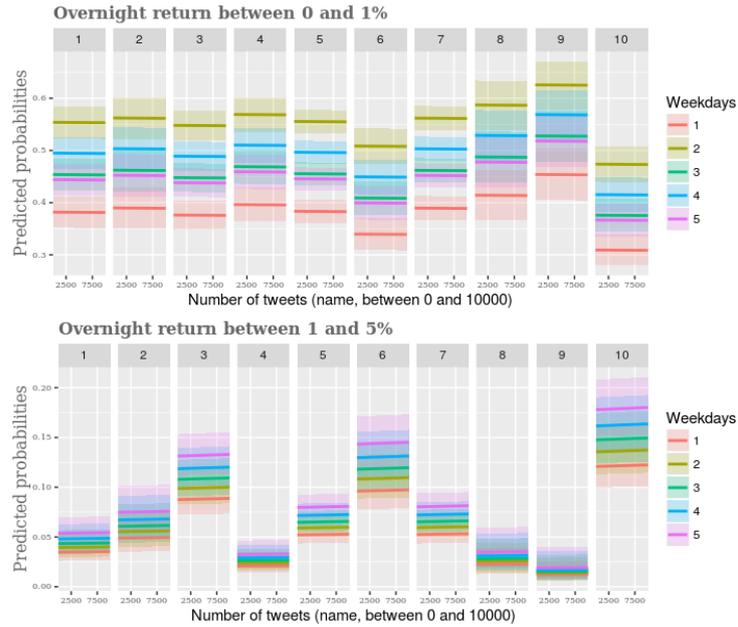


Figure 2: Predicted probabilities of having an overnight return between 0 and 1% (upper panel) and between 1 and 5% (lower panel) depending on non-financial tweets (between 0 and 10000), on weekdays (1 – Monday, 2 – Tuesday, 3 – Wednesday, 4 – Thursday and 5 – Friday) and on industry sectors (1 – Energy, 2 – Materials, 3 – Industrials, 4 – Consumer Discretionary, 5 – Consumer Staples, 6 – Health Care, 7 – Financials, 8 – Information Technology, 9 – Telecommunication Services and 10 – Utilities).

Type of tweet: name return: intraday	Model 1: 0% < Return			Model 2: 0% < Return < 1%			Model 3: 1% < Return < 5%		
	Log-odds		Odds-ratios	Log-odds		Odds-ratios	Log-odds		Odds-ratios
Constant	0.1109	*	1.1172	-0.4571	***	0.6331	-1.744	***	0.1747
Intraday_return _{t-1}	-3.247	***	0.0388	-1.061		0.3460	-2.890	**	0.0555
Name _{t-1}	-5.824 x 10 ⁻⁵		0.9999	-7.889 x 10 ⁻⁵	**	0.9999	1.861 x 10 ⁻⁶		1.0000
Weekdays (ref: Monday)									
Tuesday	0.1509	***	1.1629	0.1674	***	1.1822	2.057 x 10 ⁻²		1.0207
Wednesday	-5.226 x 10 ⁻²		0.9490	-4.403 x 10 ⁻²		0.9569	-2.026 x 10 ⁻²		0.9799
Thursday	-2.772 x 10 ⁻²		0.9726	5.157 x 10 ⁻²		1.0529	-0.1287	*	0.8792
Friday	0.3127	***	1.3671	0.2929	***	1.3403	8.539 x 10 ⁻²		1.0891
Industrial Sectors (ref: Energy)									
Materials	4.191 x 10 ⁻²		1.0428	-2.356 x 10 ⁻²		0.9767	2.117 x 10 ⁻²		1.0213
Industrials	-8.932 x 10 ⁻²		0.9145	-0.4258	***	0.6532	0.511	***	1.6668
Consumer Discretionary	1.863 x 10 ⁻²		1.0188	-8.014 x 10 ⁻²		0.9920	-6.472 x 10 ⁻²		0.9373
Consumer Staples	-0.1474	**	0.8629	-0.5137	***	0.5982	0.4628	***	1.5885
Health Care	-0.1478	*	0.8625	-0.4595	***	0.6316	0.3447	***	1.4116
Financials	-0.1648	***	0.8480	0.4351	***	0.6471	0.3726	***	1.4515
Information Technology	-1.888 x 10 ⁻²		0.9812	2.661 x 10 ⁻³		1.0026	-6.580 x 10 ⁻²		0.9363
Telecommunication Services	-0.1493		0.8612	8.709 x 10 ⁻²		1.0909	-0.5211	***	0.5938
Utilities	-0.3357	***	0.7148	-0.8585	***	0.4238	0.5359	***	1.7089
Number of observations	16043			16043			16043		

P-value levels: *** < 0.01; ** < 0.05; * < 0.1; - < 0.15

Table 7: Log-odds and odds-ratios of having an intraday return higher than 0% (Model 1), between 0 and 1% (Model 2) or between 1 and 5% (Model 3) depending on financial tweets (Ticker), weekdays and industrial sectors.

16288 tweets, Weekdays being Wednesday and Overnight_return_{t-1} being 0.00012), we can identify preferential industrial sectors for having positive overnight returns, which are similar than those previously mentioned in Table 4 (Appendix 1a). When analyzing the impact of weekdays, industry types and the number of mentions of a company simultaneously (Figure 2), the impact of non-financial tweets seems to be limited compared to financial tweets.

Intraday returns and financial tweets

As a reflection to the previous results, the impact of financial tweets on intraday returns is significant (both for Models 1 and 2, as confirmed by a Wald-test). The magnitude of Ticker_{t-1} on intraday returns is lower than on overnight returns, meaning that for an increase of one unit in Ticker_{t-1}, the log-odds of having such returns fluctuate less during trading time on the stock markets. However, the sign of Ticker_{t-1} remains the same (Table 7).

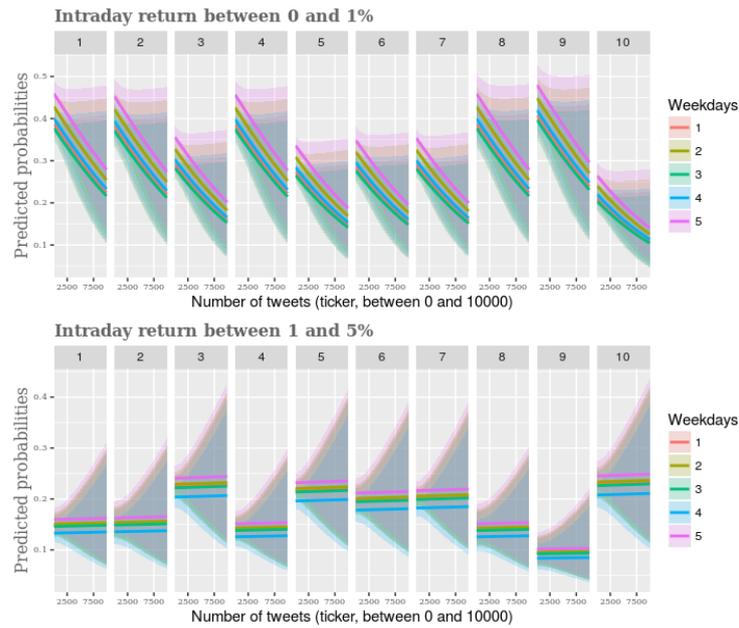


Figure 3: Predicted probabilities of having an intraday return between 0 and 1% (upper panel) and between 1 and 5% (lower panel) depending on financial tweets (between 0 and 10000), on weekdays (1 – Monday, 2 – Tuesday, 3 – Wednesday, 4 – Thursday and 5 – Friday) and on industry sectors (1 – Energy, 2 – Materials, 3 – Industrials, 4 – Consumer Discretionary, 5 – Consumer Staples, 6 – Health Care, 7 – Financials, 8 – Information Technology, 9 – Telecommunication Services and 10 – Utilities).

Type of tweet: name Type of return: intraday	Model 1: 0% < Return			Model 2: 0% < Return < 1%			Model 3: 1% < Return < 5%		
	Log-odds		Odds-ratios	Log-odds		Odds-ratios	Log-odds		Odds-ratios
Constant	0.1114	*	1.1178	-0.4553	***	0.6342	-1.746	***	0.1744
Intraday_return _{t-1}	-3.225	***	0.0397	-1.034		0.3557	-2.895	**	0.0552
Name _{t-1}	-1.859 x 10 ⁻⁷		0.9999	3.68 x 10		1.0000	8-661 x 10 ⁻⁷	*	0.9999
Weekdays (ref: Monday)									
Tuesday	0.1468	***	1.1581	0.1604	***	1.1739	2.300 x 10 ⁻²		1.0232
Wednesday	-5.752 x 10 ⁻²		0.9441	-5.289 x 10 ⁻²		0.9484	-1.724 x 10 ⁻²		0.9829
Thursday	-3.369 x 10 ⁻²		0.9668	4.188 x 10 ⁻²		1.0427	-0.1255	*	0.8820
Friday	0.3075	***	1.3600	0.2841		1.3286	8.834 x 10 ⁻²		1.0923
Industrial Sectors (ref: Energy)									
Materials	4.322 x 10 ⁻²		1.0441	-2.266 x 10 ⁻²		0.9775	2.236 x 10 ⁻²		1.0226
Industrials	-9.323 x 10 ⁻²		0.9109	-0.4316		0.6494	0.5118	***	1.6682
Consumer Discretionary	2.180 x 10 ⁻²		1.0220	-1.251 x 10 ⁻²		0.9875	-5.255 x 10 ⁻²		0.9488
Consumer Staples	-0.1445	**	0.8654	-0.5216		0.5935	0.4790	***	1.6144
Health Care	-0.1477	*	0.8627	-0.4601		0.6312	0.3458	***	1.4131
Financials	-0.1698	***	0.8438	0.4666		0.6271	0.4072	***	1.5025
Information Technology	-1.986 x 10 ⁻²		0.9803	-1.726 x 10 ⁻³		0.9982	-6.142 x 10 ⁻²		0.9404
Telecommunication Services	-0.1478		0.8625	8.928 x 10 ⁻²		1.0933	-0.5214	***	0.5936
Utilities	-0.3365	***	0.7142	-0.8594		0.4233	0.5358	***	1.7087
Number of observations	16043			16043			16043		

P-value levels: *** < 0.01; ** < 0.05; * < 0.1; - < 0.15

Table 8: Log-odds and odds-ratios of having an intraday return higher than 0% (Model 1), between 0 and 1% (Model 2) or between 1 and 5% (Model 3) depending on financial tweets (Ticker), weekdays and industrial sectors.

In Appendix 1b, we provide results for the predicted probabilities of having the three models of intraday returns while holding all variables at their means: Ticker_{t-1} equals to 127 tweets, Weekdays being Wednesday and Intraday_return_{t-1} being 0.00051. In Figure 3, the complete influence of weekdays, industry types and financial tweets could be assessed. For intraday returns between 0 and 1% (model 2), five industry types present higher probabilities of having such returns, namely the

(1) Energy, the (2) Materials, the (3) Consumer Discretionary, the (4) Information Technology and the (5) Telecommunication Services sectors. Finally, Friday, followed by Monday, are the two days of the week presenting higher intraday returns, across all industries (Figure 3).

Intraday returns and non-financial tweets

In Table 8, we compute the log-odds and odds-ratios for each model and found that the impact of Name_{t-1} is not significant

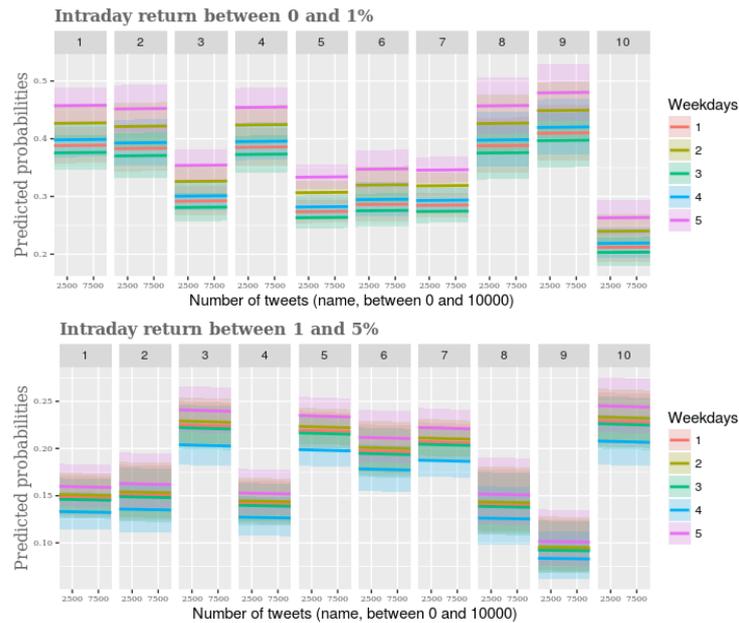


Figure 4: Predicted probabilities of having an intraday return between 0 and 1% (upper panel) and between 1 and 5% (lower panel) depending on non-financial tweets (between 0 and 10000), on weekdays (1 – Monday, 2 – Tuesday, 3 – Wednesday, 4 – Thursday and 5 – Friday) and on industry sectors (1 – Energy, 2 – Materials, 3 – Industrials, 4 – Consumer Discretionary, 5 – Consumer Staples, 6 – Health Care, 7 – Financials, 8 – Information Technology, 9 – Telecommunication Services and 10 – Utilities).

for positive intraday returns (Model 1) and between 0 and 1% (Model 2). Again, the influence of non-financial tweets is lower than financial tweets, as illustrated by the amplitude of the log-odds for each type of variables.

After computing the predicted probabilities of having the three returns while holding all variables at their means (Appendix 1c), Figure 4 confirms that using non-financial tweets for predicting such returns is not effective.

CONCLUSION

Two main findings of this study have been established. Firstly, financial tweets are more relevant to take into account compared to non-financial tweets regarding intraday and overnight returns. While the effect of the first ones does impact negatively on lower returns (0 to 1%) and positively on higher returns (1 to 5%), non-financial tweets did not help in explaining these returns. Secondly, overnight returns are more impacted by financial tweets than intraday returns, confirming our first hypothesis: tweets contain information that seems to be reflected on stock markets as soon as they open on the next day, which is less capture during trading hours. From these results, investment dashboard could be built, controlling for industry types, weekdays and live-monitoring of Twitter data.

Being exploratory, our research has some limitations and further work could be done in order to improve the quality of our findings. Firstly, we only selected firms that are already popular on Twitter (or at least having 30 financial messages per day on average). Secondly, the impact of a single tweet is low compared to traditional variables (returns of the previous day for example). This can be interpreted as a result that Twitter messages cannot fully explain returns on the stock market. However, these unstructured data can add value to traditional investment strategies based on more regular sources of information. In this regard, Twitter can add some information, being an evolution, but maybe not a revolution in finance.

In terms of estimation techniques, we think further work should be done, for instance using a straight OLS while identifying and correcting potential problems with unstructured data such as potential non-linearity of the interacted variables, weak extrapolation and severe interpolation. We could also implement machine learning techniques such as neural network or cluster-based analyses.

REFERENCES

- [1] Markowitz H, Portfolio Selection, The Journal of Finance, vol. 7, no. 1, pp. 77–91, 1952.
- [2] Fama EF, Efficient Capital Markets: A review of theory and empirical work, The Journal of Finance, vol. 25, no. 2, pp. 383–417, 1970.
- [3] Nofsinger JR, Social mood and financial economics, The Journal of Behavioral Finance, vol. 6, no. 3, pp. 144–160, 2005.
- [4] Akerlof GA and Shiller RJ, Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism, Etats-Unis: Princeton University Press, 2009.
- [5] Tapiero CS, The Future of Financial Engineering (SSRN Scholarly Paper No. ID 2259232). Social Science Research Network, Rochester, NY.
- [6] Warin T and Sanger W, Structuring Big Data: How Financial Models May Help, Journal of Computer Science and Technology, vol. 2, no. 1, pp. 1–20, 2014.
- [7] Warin T, De Marcellis-Warin N, Troadec A, Sanger W and Nembot B, Un etat des lieux sur les données massives (No. 2014RB-01), Montréal, Canada: CIRANO, 2014.
- [8] Tetlock PC, Saar-Tsechansky M and Macskassy S, More Than Words: Quantifying Language to Measure Firms' Fundamentals, The Journal of Finance, vol. 63, no. 3, pp. 1437–1467, 2008.
- [9] Wysocki P, Cheap Talk on the Web: The Determinants of

- Postings on Stock Message Boards by Peter D. Wysocki, SSRN, Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia, 1998.
- [10] Tumarkin R and Whitelaw R, News or Noise? Internet Postings and Stock Prices, *Financial Analysts Journal*, pp. 41–50, 2001.
- [11] Antweiler W and Frank MZ, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.
- [12] De Choudhury M, Sundaram H, John A and Seligmann DD, Can blog communication dynamics be correlated with stock market activity? In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, New York, 2008.
- [13] Gruhl D, Guha R, Kumar R, Novak J and Tomkins A, The predictive power of online chatter. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, New York, USA, 2005.
- [14] Buechel B, Hellmann T and Kloessner S, Opinion Dynamics under Conformity (Working Paper No. 469), Center for Mathematical Economics, Bielefeld University, 2012.
- [15] Brink RVD, Rusinowska A, Steffen F, Measuring Power and Satisfaction in Societies with Opinion Leaders: An Axiomatization, Université Paris1 Panthéon-Sorbonne, 2011.
- [16] Song X, Chi Y, Hino K, and Tseng B, Identifying opinion leaders in the blogosphere, in proceedings of the sixteenth acm conference on conference on information and knowledge management, New York, NY, USA, ACM, pp. 971–974, 2007.
- [17] Domingos P, and Richardson M, Mining the network value of customers in proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, New York, NY, USA, ACM, pp. 57–66, 2001.
- [18] Kempe D, Kleinberg J, and Tardos E, Maximizing the spread of influence through a social network in proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, New York, NY, USA ACM, pp. 137–146, 2003.
- [19] Nakajima S, Tatemura J, Hino Y, Hara Y, and Tanaka K, Discovering important bloggers based on analyzing blog threads, 2005.
- [20] Varian HR, Beyond Big Data, *Business economics*, vol. 49, no. 1, pp. 27–31, 2014.
- [21] Choi H, and Varian H, Predicting the present with google trends, *Economic record*, vol. 88, pp. 2–9, 2012.
- [22] Preis T, Reith D, and Stanley E, Complex dynamics of our economic life on different scales: insights from search engine query data explore research WBS, *Philosophical Transactions of the Royal Society*, pp. 5707–5719, 2010.
- [23] Dimpfl T, and Jank S, Can internet search queries help to predict stock market volatility? university of cologne, centre for financial research (CFR), no. 11-15, 2011.
- [24] Bank M, Larch M, and Peter G, Google search volume and its influence on liquidity and returns of german stocks, Rochester, NY: Social Science Research Network, 2010.
- [25] Bordino I, Battiston S, Caldarelli G, Cristelli M, Ukkonen A, and Weber I, web search queries can predict stock market, vol. 7, no. 7, 2012.
- [26] Da Z, Engelberg J, and Gao P, In search of fundamentals Rochester, NY: Social Science Research Network, 2011.
- [27] Ramos SB, Veiga H, and Latoeiro P, Predictability of stock market activity using google search queries (statistics and econometrics working paper No. ws130605), Universidad Carlos III, Departamento de Estadística y Econometría, 2013.
- [28] Zhang W, Shen D, Zhang Y, and Xiong X, Open source information, investor attention, and asset pricing, *Economic modelling*, vol. 31, pp. 613–619, 2013.
- [29] Mao H, Counts S, and Bollen J, Predicting financial markets: comparing survey, news, twitter and search engine data, Vol. 14, 2011.
- [30] Karabulut Y, Can facebook predict stock market activity? (ssrn scholarly paper no. id 1919008), Rochester, NY: Social Science Research Network, 2011.
- [31] Kramer AD, Guillory JE, and Hancock JT, Experimental evidence of massive-scale emotional contagion through social networks, *Proceedings of the national academy of sciences*, 2014.
- [32] Pan W, Altshuler Y, and Pentland A, Decoding social influence and the wisdom of the crowd in financial trading network, pp. 203–209, 2012.
- [33] Altshuler Y, Pan W, and Pentland A, Trends prediction using social diffusion models, In Yang SJ, Greenberg AM, and Endsley M (Edition), *social computing, behavioral - cultural modeling and prediction*, Springer berlin heidelberg, pp. 97–104, 2012.
- [34] Pan W, Cebrian M, Dong W, Kim T, Fowler J, and Pentland A, Modeling dynamical influence in human interaction patterns, Vol. 14, 2010.
- [35] Bollen J, Mao H, and Zeng, X-J, Twitter mood predicts the stock market, *Journal of computational science*, pp. 1–8, 2011.
- [36] Brown E, Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market, SAIS 2012 Proceedings, 2012.
- [37] Sprenger T, and Welpel I, Tweets and trades: the information content of stock microblogs (ssrn scholarly paper no. id 1702854), rochester, NY: Social Science Research Network, 2010.
- [38] Porshnev A, Redkin I, and Shevchenko A, Improving prediction of stock market indices by analyzing the psychological states of twitter users (SSRN scholarly paper no. id 2368151), Rochester, NY: Social Science Research Network, 2013.
- [39] Warin T, Marcellis-Warin, de N, Sanger W, Nembot B, and Mirza VH, Corporate reputation and social media: a game theory approach, *international journal of economics and business research*, Vol. 9, no. 1, pp.1-25, 2015.
- [40] Leavitt A, Burchard E, Fisher D, and Gilbert S, The influentials: new approaches for analyzing influence on twitter, 2009.
- [41] Vincent A, and Armstrong M, Predicting break-points in trading strategies with twitter (SSRN scholarly paper no. id 1685150), Rochester, NY: Social Science Research, 2010.
- [42] Zhang W, and Skiena S, Trading strategies to exploit news sentiment, proceedings to the fourth international AAAI conference on weblogs and social media, 2010.
- [43] Mackintosh J, Last tweet for Derwent’s Absolute Return, *Financial Times*, 2012.